Acoustic Convergence: Exploring the Influence of Ambient Noise on Speech Production

Ryan Gary Podlubny

Department of Linguistics

University of Canterbury

This dissertation is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Linguistics

September 2019

Abstract

Increasing work across disciplines has recognized various ways in which people can be influenced by aspects of their surroundings. In fact, people have even been found at times to take on certain properties of the environment or objects within it. This process of *convergence*, where people align in particular ways with characteristics of their surroundings, has also been noted in a variety of linguistic-research and is the focus of the present work. While acoustic-convergence between speech partners is well attested at this point, little is known about how speech production may be influenced by ambient noise in the speech environment. Aiming to address this gap in the literature, the present work outlines the execution and analyses of three speech-in-noise experiments designed to test whether or not productions become systematically more (and/or less) like ambient noise in the dimensions of pitch, intensity, and speech rate/tempo. Furthermore, this work also investigates potential drives for any such convergent/divergent behaviours, as rooted in primarily automatic, social, and experiential motivations.

Two forms of noise are explored in the present work. Specifically, background *music* appears in each of the three studies, and background *speech* is introduced during Experiment 3 in a way that allows for direct comparison of convergence effects across noise types. A number of experience- and socially-based variables are tested for potential contributions when convergent/divergent behaviours are shaped.

This dissertation provides evidence that speech production is affected in reliable ways by background music, such that speech converges and diverges acoustically with aspects of that noise. A self-reported estimate of music listened to per day and a speaker's identified gender are recognized as the most powerful predictors across experiments. As argued by Babel (2009), neither a wholly automatic nor socially-based theory of convergence can sufficiently account for the present data, and it appears clear that both play a role in shaping these changes to speech production. The effects observed in this work closely resemble convergence between speakers in previous sociolinguistic research, where personal attitudes and alliances are reflected through either convergence or divergence – moreover, certain forms of personal experience are recognized to further mediate these effects. The present work therefore provides valuable results for future work investigating theories of speech perception/production, and suggests that we as speakers are constantly updating our production patterns to become more (or less) like our surroundings.

Acknowledgements

The road to this moment has been long and interesting – fun at times, less fun at others – but valuable experience wholly. There is no way I could have kept things from going off the rails through this degree without a great deal of help, and I would like to take a minute to send love to those who have given so much to me throughout this process.

First and foremost, I cannot say enough good things about Jen Hay. I could not have hoped for anything more in a supervisor: Your patience, enthusiasm, respect, guidance, and boundless talent have been an absolute pleasure to experience over the last four years, and I feel extremely fortunate to have you as a colleague and as a friend. Vica Papp has also shown me unending patience and skill as my friend and secondary supervisor: you've given me much more elegant ways to write code, some interesting art-house Hungarian cinema, and offered many an insight through fun and stimulating conversation. There is no better vegan mac-n-cheese.

I would to send a general 'thank you' to everyone down the NZILBB/Linguistics corridor on the 2nd floor of the Locke building for helping me to flesh out ideas, for pizza and beer breaks, and for looking at photos of LoPan on my phone. Specifically, I owe thanks to Clay Beckner & Petya Rácz for putting up with my constant questions about statistics, for many a good conversation, and for the occasional jam session. I received HUGE amounts of support from my fellow students throughout my time at Canterbury: Darcy Rose, Ksenia Gnevsheva, Matthias Heyne, Keyi Sun, Andy Gibson, Xuan Wang, Daniel Bürkle, Jacq Jones, Mineko Shirakawa, Vicky Watson, Sidney Wong, and Sara van Eyndhoven – you're the best. Of course there was also much encouragement from Kevin Watson, Heidi Quinn, Donald Derek, Lynn Clark & Sarah Hawkins: thank you all for treating me so well, and for helping me grow as a scholar. I would like to thank Jacqui Nokes, Marton Soskuthy, Katie Drager (thanks again for taking me to Costco!), Robert Fromont, Christopher Dromey, Deryk Beal, Donal Sinex, Daniel Gerhard, Francois Bissey, Fabian Tomaschek and Ben Tucker for allowing me to bounce ideas off of you, and for taking the time to help me improve them. Of course I also owe a great debt to Emma Parnell for both professional help and listening to me complain a lot – you're as good as they come!

Dave Bui, Astrid Simonsen, Max Capocaccia, and Alison Cook: thank you all for so many wonderful nights of music, food, relaxing, and time focused on [-linguistics]. Danielle, Roan, Saoirse and Clover of team Ferreira/Beckner: Family dinners and all the dog parties are some of my favourite memories in New Zealand – I don't have words to express how much I appreciate you welcoming me into your family. Movie nights with Greg Baker came as a welcome reprieve from my studies, and the pierogi parties, road trips, and dog-park-days with Anastasia Yuchshenko more or less kept me sane. Then there are those who either installed me with neuro-skeletal upgrades, or kept me alive while healing after the fact: I'd like to send enormous thanks to Kris Dalzell and Michelle Angus, Gabby Watson, Julia and Kai Zimmerman, Satellite, Kathryn Shaw, Bob Hay, Josephine Varghese, Andrew Kepple, Ni Zhou, Annalise Fletcher, Jamie Cho, Becci Neal, Molly Watson, and of course my live-in Columbian (and bringer of cottage cheese) Fernando Cagua. I would also like to thank Rob Batke for the time and love you put into co-composing the music used as stimuli within this work (Science Music), and of course all speakers who took part in these studies.

Ken, Linda and Greg Podlubny, Lauren and the family Hawes, and Andrea Tam: Thank you for years of support, guidance, and patience. You have all done so much to make this goal achievable for me, and this book could not exist without you. Thank you to Grant McGuire and Rebecca Scarborough for their many thoughtful comments during the examination process. Thank you to Lisa Haynes for helping me limp across the finish line – and finally, thanks to LoPan for reminding me to take breaks (even if it was only so I could rub her belly).

This research was funded in part by the Social Sciences & Humanities Research Council of Canada (752-2014-1438-SSHRC) and the University of Canterbury's doctoral scholarship.

TABLE OF CONTENTS:

LIST OF FIGURES:

LIST OF TABLES:

CHAPTER 1: Introduction and Literature Review

1.1 Introduction

As humans making our way through the world, we begin to learn from a very young age how to communicate with others. Of course this process often begins with gaining proficiency in our initial language(s), beginning with rudimentary phonological awareness and moving on to small words and phrases. Eventually, we gain strong intuitions that coincide largely with the experience and expectations of those who make up our surrounding speech community. However, attaining native proficiency in a language (or languages) does not mark the end of our linguistic development.

Language users continue to adopt new lexical items; learn to generate and understand increasingly complex syntactic structures; and, of course, the ways in which we produce speech sounds over time also continue to change as we develop. Such instability is in part due to human physiological changes over time, namely in the vocal anatomy (e.g., Ryan & Burk, 1974; Schötz, 2007: pp. 88-90) – though, can also occur as the product of learning and/or personal experience.

Indeed, previous works have shown that even relatively stable speech production patterns can be influenced and altered to some degree through a speaker's social identity and opinions (Babel, 2009: pp. 143-144; Drager, Hay & Walker, 2010), interactions with other speakers (e.g., Pardo, Gibbons, Suppes & Krauss, 2012; Delvaux & Soquet, 2007a), and some changes have even been noted as a result of exposure to ambient noise through the so-called Lombard effect (e.g., Lombard, 1911; Lane & Tranel, 1971). Thus, it is currently well accepted that speech production can be influenced in various ways by other talkers, and by the speaker's attitudes.

It has also been shown that speech production is affected differently by different types of background noise, as well as by physical properties of the speech environment (e.g., Junqua, 1996; Stowe & Golub, 2013). Some work has even shown that speech can be shaped by previous experience and expectation in a given situation (Hay, Podlubny, Drager & McAuliffe, 2017). However, while the literature on speech production continues to grow with respect to how patterns may change as a function of context, the specifics of how speech is influenced by ambient noise are largely unknown at this point outside of work on the better-known Lombard effect. As the current best-understood phenomenon related to ambient noise influencing acoustic characteristics of speech production, the Lombard effect will be discussed in detail during the following literature review (Section 1.2) as this stream of research served to guide many of the decisions involved in designing the experiments which comprise this work.

This dissertation, therefore, recognizes an important gap in the literature insofar as certain ways that background noise seems likely to affect speech production appear relatively understudied. Where it is known through a sizable literature on *speech accommodation and convergence* that human speakers often alter patterns of their spoken language, both consciously and unconsciously, to become more – or sometimes *less* – like those of other speakers or speech partners (e.g., Shockley, Sabadini & Fowler, 2004; Goldinger, 1998: Luce, Goldinger, Auer &

Vitevitch, 2000; Giles, 1973; Giles, Mulac, Bradac & Johnson, 1987; Giles & Powesland, 1997: pp. 232-239; Street, 1984; Platt & Weber, 1984; Pardo, Urmanche, Wilman, & Wiener, 2017; Babel, 2009; Babel, 2010; Hay, Drager & Warren, 2009; Drager, Hay & Walker, 2010), it is currently unknown whether or not speakers' productions similarly gravitate toward (*and/or away from*) acoustic characteristics of ambient noise. There are no known studies that test for this type of acoustic convergence to non-speech background noise. In fact, only two known articles from Delvaux & Soquet (2007a; 2007b) directly explore the possibility of acoustic convergence to backgrounded *speech* – that is, spoken language experienced by the talker where there is no explicit interaction with an interlocutor, nor is the background speech related in any direct way to the experimental task (cf. *Shadowing* e.g., Marslen-Wilson, 1973; Goldinger, 1998).

Knowing through previous work that speakers often converge acoustically with other speakers, testing for entrainment in speech production to ambient noise contributes meaningfully to a number of theoretical and practical streams of research. In the way of theory, providing evidence that speakers converge/diverge with background noise would suggest that human speech production patterns are constantly changing, and being influenced by whatever sounds exist around us. If speakers are in fact shown to be sensitive to the speech environment in this way, then the present work could influence the design of future phonetic enquiry as well as the interpretation of some previous studies – put simply, if speakers are shown to constantly alter production patterns to be more (or less) like ambient noise, then speech scientists must keep this sensitivity in mind when designing experiments as it would almost surely influence some experimental outcomes. In a more practical sphere, this research could also lend itself to improving various communicative technologies: The studies that follow contribute to better understanding how human speakers deal with noisy environments during speech production, and specifically how speech may change in predictable ways due to ambient noise. Human listeners are normally very good at focusing on specific talkers despite noisy environments (i.e., selective attention and filtering), though modern technologies have not yet matched this ability. Therefore, identifying potential (regular) changes in speech patterns driven by noise could lead to the development of more effective environmental filtering algorithms.

However, negative results would also be informative. It could have been the case that no convergence/divergence was observed in the present work, or effects may have been observed in varying degrees as a function of noise type – which would speak directly to the present knowledge of speech accommodation. For example, if convergence was observed in only the speech-based conditions (and not to background music), this result would suggest convergence as a primarily social process, whereby speakers update productions to include indexical information that serves to externalize affiliations (or a lack thereof). However, it was also possible that speakers would not exhibit convergent behaviours in either noise-condition which, in light of works from Delvaux & Soquet (2007a, 2007b), would have suggested that speakers converge to ambient noise on a phonemic level, but not to speaker prosody. If this were the case then the social-nature of convergence would have still been supported, though the extent to which speakers/listeners are sensitive to any such alignment may not sink below the level of linguistically meaningful units. Importantly, the experiments that follow do, in fact, indicate that speakers alter their speech to become more/less like ambient noise, and I will explain throughout this dissertation how I believe the reasons for this convergence and divergence parallel certain effects related to convergence in communicative speech.

In other words, a social drive for convergence and divergence is well supported through the experiments that follow, where altered speech patterns appear to reflect speaker attitudes and affiliations.

Convergence in speech production is often regarded as social in nature (e.g., Drager, 2011), though some theories presume absolute automaticity in the mechanisms that drive this phenomenon (e.g., Trudgill, 2008). Notably, work from Babel (2009) indicates that convergence in speech production involves both automatic and social components, and that neither a purely automatic nor a completely social theory based on conscious mechanisms for accommodation was supported by her work. Opposing theories driving convergence are discussed in the literature review below (section 1.2.6), and how such theories are impacted by the present findings will also be discussed throughout the following pages.

For the sake of clarity, I will take a moment to draw an important distinction between 'socially-based' effects as they relate to linguistic convergence, and 'experience-based' effects. I will use the term *socially-based* effects when referring specifically to convergence-related phenomena that involve the process of aligning and/or distancing oneself with a conversation partner (or similar) or expressing speaker attitudes through a variety of linguistic means e.g., sentence structure, speech patterns, lexical choices, etc. – these processes are therefore recognized for having social motivations. Conversely, I will use the term *experience-based* effects when discussing altered speech patterns driven by learned or conditioned mechanisms that do not serve primarily to align or distance the speaker in any socially-specific context. Experience-based effects may often contribute to improved intelligibility, which, in a sense may appear social in nature because communicative acts tend to be rather social events. However, effective message transmission does not necessarily, in and of itself, align speaker and hearer; whereas socially-based effects must involve some sort of stratification, relative alignment or distancing between speakers or speech communities, or otherwise implicit social motivation (even if abstracted cf. Drager, Hay & Walker, 2010), thus distinguishing these two concepts.

But in much the same way that talkers often converge in their speech patterns, there is sufficient motivation to expect that speakers may *also* alter productions to become more like ambient noise – such motivations are described in detail in the following literature review. Therefore, through a series of three experiments the present work has been designed to explore the influence of both background music and background speech on human speech production, addressing the above-mentioned gap in the literature. The work that follows will be presented thusly.

Chapter one serves to introduce the phenomenon of convergence in speech production. Previous works are discussed, motivating the research questions addressed through this dissertation. Specifically, a general tendency toward convergence is recognized as ubiquitous in nature both inside and outside the scope of linguistics. It is proposed that because human subjects are known to converge in a variety of linguistic and non-linguistic ways, that speakers' productions are likely to converge acoustically with aspects of ambient noise as well. Research questions are stated which explore the potential for convergent and divergent behaviours in an acoustic context as random or systematic processes, and an overview of the three experiments described within this collection is provided.

The second chapter of this work describes Experiment 1, a speech-in-noise reading task designed to address the questions posed in section 1.3 through exposure to carefully constructed and manipulated background music. As the first (known) experiment to investigate the possibility of convergence to background noise, much of this chapter

is dedicated to developing methodologies that appropriately and effectively test the questions posed. This discussion involves in-depth description of numerous concepts and characteristics related to music that may influence participants in unintended ways, as well as how such issues were mediated. These considerations are broken up into two groups, being *general* (that is, primarily rooted in psycholinguistic associations given previous experience) and the *specific* (being related more so to physical/acoustic properties of the signals themselves, and how these properties may affect processing). This study explores the possibility of convergence in voice-pitch, talker intensity, and in speech rate. Generation of the acoustic manipulations used to test for such convergence is also discussed at length due to the introduction of novel software previously unused in phonetic experimentation. Use of a proprietary platform designed for use by musicians and DJs (i.e., Ableton's Live 9) resulted in good quality signal manipulations – however, little is known about how the software achieves these manipulations. As a result, extensive post-manipulation analyses were run and are described to ensure only the intended acoustic characteristics had been altered and, thus, test the efficacy of these methods. Having confirmed high-quality stimuli had been generated in the intended way(s), results from this experiment provide some evidence for a systematic influence of background noise, though the precise mechanisms driving these effects are not yet clear. Speakers' previous musical training is observed as the best available predictor for convergent and divergent behaviour to pitch and intensity variation. No effects are observed for speech rate. However, due to a low-powered analysis stemming from weak participant numbers and issues with the data, means to improve the design in a replication study are considered.

Chapter 3 describes Experiment 2 (EXP.2), a re-designed speech-in-noise production task aiming to (1) replicate effects observed in EXP.1, and (2) address both analytical and design issues that surfaced during the first experiment. This study recognizes that investigating convergence to tempo is an extremely complicated problem and therefore narrows the scope of EXP.2 to exploring voice-pitch and vocal-intensity only, reserving further investigation of convergence to tempo for future work. In order to address certain issues encountered during the analysis of EXP.1, the task and stimuli have been changed somewhat to facilitate more straightforward statistical modeling. Another primary change to this study involves an altered approach to the intensity-based manipulation. Specifically, the presentation level of noise is implicated in the Lombard literature as providing a threshold for effects observed – Lombard speech is elicited only through noise of at least a certain perceived loudness, and this threshold changes based on whether the background signals are comprised of speech or non-speech signals. The altered intensity manipulation in this study dips below the known threshold for eliciting the Lombard reflex in hopes of distinguishing a Lombard response from entrainment to signal intensity. New, more extensive analytical methods are introduced, where results suggest that the amount of time a speaker spends listening to music per day is in fact a more reliable predictor for convergent and divergent behaviours in speech production (though, musical training is still observed to contribute in a relatively limited context). Crucially, evidence is provided for convergence to intensity below the known threshold for eliciting Lombard speech, indicating either (1) that entrainment to intensity is in fact distinct from the Lombard effect, or (2) that the threshold for eliciting Lombard speech is lower than previously understood. Given the divergent behaviour observed from participants who listen to relatively little music per day and the convergence observed below this known threshold, a distinction between Lombard and convergence appears more likely. Thus, hypotheses with regard to the mechanisms driving convergence/divergence to ambient

noise become more focused, and suggest the possibility of effects rooted in *social behaviours* – that is, convergence and divergence appear to reflect speaker attitudes and opinions as a function of listening habits.

Chapter 4 describes EXP.3, a study in which the investigative scope narrows further to exploring convergence in voice-pitch only. Assumptions were made throughout the preceding experiments based on the only known work to explore convergence to ambient speech (Delvaux and Soquet, 2007a; 2007b); however, potential issues within these works are identified and discussed, focussing on how results in these studies may have been misinterpreted. Consequently, new experimental conditions are introduced within EXP.3 to investigate the potential for convergence to ambient speech as well as the backgrounded music (the design of this study is otherwise identical to EXP.2, aiming to later jointly analyze data across experiments). Exactly how different acoustic-convergence to speech may be from convergence to background music is currently unknown. To address this issue, conditions are generated such that any effects observed are directly comparable in analysis across conditions. Much like in EXP.2, results from random forests are used to feed linear mixed effects models. Musical training is again implicated as a predictor for convergence and divergence – though, in this case convergence to background speech. Whether or not a speaker chooses to listen to music while undertaking cognitively demanding tasks is found to best predict convergence/divergence to background music in these data, suggesting that a speaker's musical consumption/listening habits are in fact driving convergence to background music. Importantly, speaker listening-habits interact with identified gender in a way that further influences performance significantly. Therefore, a single, clear explanation for speaker-convergence is not yet available through a single variable, as predictors retained in modeling across experiments differ; however, a shared drive for these effects is implied through thematically overlapping results across models, suggesting the observed effects may be the result of *musical preferences* (i.e., a socially-based impetus) with further influence of personal experience. The influence of listening habits is tested through a unified analysis in the following chapter.

At this point in the cumulative work, results across studies appear relatively congruent when considered together – though, more robust analyses may be gained by combining data across studies. To this aim, I attempt a unified analysis in chapter 5, incorporating data from all three studies (data were transformed to make them relatively more comparable across studies within a single analysis). Results support recurring themes observed in the individual analyses, confirming effects observed in previous analyses. Speaker convergence and divergence are further supported as predictable through both a speaker's identified gender and through musical consumption habits.

The sixth, and final chapter serves as a general discussion. Experimental findings are summarized, and discussed in light of each other. The research questions posed in Chapter 1 are addressed directly while considering the preceding unified and individual analyses. A clear motivation for convergence and divergence in speech production appears to be available through speaker listening habits, which seems to be further shaped by a speaker's identified gender. It appears that convergence and divergence to background music is somewhat like that to communicative speech – that is, speakers are observed to converge with things they like (as evidenced though speakers with higher levels of musical consumption per day converging) and speakers tend to diverge from things they dislike (as seen in speakers who listen to relatively less music exhibiting divergent tendencies). The present collection as a whole, therefore, supports both social- and experiential drives – both with some degree of

automaticity – for convergence/divergence, as well as a need for further study of acoustic convergence to ambient noise. Given the repeated trends, the observed effects across studies appear to be reliable, and not just random patterning in the data. However, before diving into the details of these experiments, it makes sense to first explore previous experimentation that has led to the present enquiry. Indeed, a discussion of convergence generally is in order, and can be found directly below in section 1.2.1.

1.2 Background

1.2.1 What is Convergence?

Through a variety of studies, linguistic and otherwise, a general trend has emerged in nature where various forms of synchronization to external sources are found in the environment. Human *convergence* is being studied increasingly across disciplines, many of which have coined or adopted a number of terms (e.g., *entrainment, alignment, convergence, imitation, synchrony, mimicry,* and/or *accommodation*) to describe the types of synchronization known to influence human motor control, neurology, and even human physiology (e.g., Hill, Adams, Parker & Rochester, 1988; Will & Berg, 2007; McClintock, 1971). It appears entrainment is a rather common occurrence in many domains for human subjects, and is something that most often occurs both automatically and unknowingly.

There is reasonable evidence supporting that human subjects at times align with both linguistic and non-linguistic aspects of their environment (detailed below) – and through these streams of research, it is further known that we as humans experience forms of convergence and divergence both in- and outside of a communicative context. Importantly, 'convergence' is typically defined as *a process whereby specific characteristics of two (or more) distinct entities become more alike over time through contact*. However, due to the nature of the present work, this definition must be altered somewhat. Specifically, where convergence typically involves at least two entities that might influence each other, changes in the background signals used in each of the following studies are (necessarily) independent of participants' speech productions. Simply, background music cannot be influenced by a speaker's productions in this context, though a speaker's productions *may* be influenced by the music. This relationship is therefore inherently unidirectional, unlike most contexts where convergence might be observed. Another issue is raised when considering how exactly one might converge to background music, where any observed changes could be *absolute* or *relative* in nature. For example, it is possible that a speaker's voice-pitch might shift to become nearer the pitch of the music (an absolute change involving a specific starting point and a specific target), and it is also possible that a talker's voice-pitch could follow the time-varying pitch-envelope of that background signal (a relative change, whereby patterns are adopted instead of specific targets). Because both background music and multi-talker babble are complex signals, each involving multiple independent voices that simultaneously exhibit various streams of pitch, rhythm, etc., specific acoustic-phonetic targets cannot be identified as any such targets are likely to differ by listener (cf. Auditory Scene Analysis: Bregman, 1994; more on this below). Therefore, context limits the present work to testing for relative changes in production. As a result, I extend the above definition to include *situations*

where movement observed for one entity in a particular acoustic dimension influences movement in another entity, such that the two exhibit similar change over time as forms of convergence. Note that throughout this work I refer to convergence-based studies and theory that were not developed with my relatively broader conceptualization of this process in mind; however, I am arguing that convergence need not be restricted to absolute targets, but can also include aligning with patterns. With this argument in mind, the studies cited throughout this dissertation have been applied in line with my conceptualization of convergence.

In the review below I discuss different forms of what would more typically be referred to as convergence, beginning with a brief summary of some findings outside of human study. I next discuss findings related to physical convergence in human subjects, and move on to various forms of linguistic convergence attested in the literature. I provide a description of mechanisms argued to drive convergence, which precedes a summary of the Lombard literature (which is much like a form of acoustic entrainment, at least in certain ways). This discussion also includes some relevant information from the speech-in-noise literature, which also served to inform the studies described herein.

1.2.2 Convergence in Non-Humans

Now that we have a better idea of what is meant by convergence, I would like to emphasize that this phenomenon is not a rare occurrence in the natural world. Before I focus on convergence-based inquiries which are directly related to human behaviour, I will first briefly summarize a sample of experiments which suggest convergence to external influence is, in fact, rather pervasive in nature, and broad in how different forms of synchrony are realized.

One of the more popular examples of non-human convergence is categorized as *collective synchronization,* and is often used as a demonstration for coupling oscillators in Physics classrooms. The earliest known discussion of this type of convergence dates back to 1657 and was noted through Christian Huygen's work refining pendulum clocks (Hugenii, 1673). In this context, the synchronization involves a "system of oscillators spontaneously lock[ing] to a common frequency, despite the inevitable differences in the natural frequencies of the individual oscillators" (Strogatz, 2000). Put simply, the system of oscillators – such as a group of metronomes started in various degrees of out-of-phase on a freely moving platform – are loosely coupled as a result of the play in that platform, and so will gradually but eventually synchronize with each other at a single frequency due to increasingly shared movement when given sufficient time (e.g., Pantaleone, 2002). These and similar works introduced a situation where multiple, inanimate objects automatically synchronize in physical movements, noting no potential contribution of agency.

Other forms of convergence have been noted in living, non-human organisms as well; however, these forms of convergence are typically regarded as evolutionary in drive. Studies from various biologically-based disciplines show that, for example, certain species of periodical cicadas regionally synchronize lifecycles naturally and automatically, and emerge together every 13 or 17 years dependent on the sub-species (Lloyd & Dybas, 1966). Similar work from Mirollo and Strogatz (1990) describes how fireflies congregate in trees in certain areas in Southeast Asia and begin to flash in synchrony. Indeed, much like the behaviours observed in cicadas and fireflies, Walker (1969) describes convergence observed in the chirps of the Snowy Tree Cricket. These insects alter the phase

differences of their chirps based upon the timing of neighbouring chirps, resulting in the orchestra (the group of crickets) eventually chirping in unison. Beyond insects, convergence has also been noted in certain species of elephants, where the reproductive cycles of females of lower status have been found to converge with those of the dominant female (Weissenböck, Schwammer & Ruf, 2009).

The social aspect involved in convergence noted here is of the utmost importance to the present work, as the drive for convergence as something rooted in automaticity vs. a social motivation has been long discussed in human-focused convergence studies (this topic is addressed later, in section 1.2.5). Therefore, one final non-human example is provided below to reinforce the importance of social factors as they may relate to convergence-based effects generally. This example comes from a study exploring contact vocalizations in young goats. Briefer & McElligott (2012) describe vocal plasticity observed in young kids while investigating whether or not calls are affected by social environment and kinship. The authors compared calls from half siblings raised in the same group to those raised in different groups, finding that calls of half siblings "…were more similar when they had been raised in the same social group than in different groups, and converged with time." This type of work reinforces the influence of social interaction on vocalization patterns outside of human language use, and further supports some forms of convergence generally involving social motivations.

Thus, various forms of convergence are well attested in nature, and have been observed in a broad range of realizations spanning physiological to communicative. Next, I explore convergence as it has been observed in human-based research.

1.2.3 Physical Convergence in Humans

As noted above in 1.2.1, varied forms of non-linguistic convergence have been observed in human subjects as well. For example, while the phenomenon has since been argued to be more complex than first thought (e.g., Pettit & Vigor, 2015), studies from McClintock (1971); Russel, Switz, & Thompson (1980); Stern & McClintock (1998); and from Graham and McGrew (1980) describe estrous (menstrual) synchrony occurring naturally in human females, not unlike the study from Weissenböck et al. (2009) mentioned above. These papers are particularly interesting when considered together however. While some degree of estrous synchrony is noted in each of the above studies, the importance of both pheromones and of social relationships has been noted in different research on this form of synchrony. That is to say, both automatic/biological factors and social/interactional factors have been implicated as contributing to this type of convergence and how it is realized. With this in mind, I would like to draw attention to the importance of context as another crucial and recurrent theme throughout the present work, be it rooted in personal experience or more physical properties of the subject or testing environment. The greater context in which we experience events will come up in the chapters that follow as an important contributor to our roles in those events, as well as our interpretations of them.

Beyond reproductive biology, we also encounter various forms of automatic convergence in human neurology. For example, EEG-based work from Galambos, Makeig, & Talmachoff (1981) and from Will & Berg

(2007) describes how human brainwaves are known to converge with auditory stimuli. Will & Berg requested participants listen to auditory stimuli passively while EEGs were recorded. The study focussed on periodic signals comprised of "clicks" and "drum sounds", though listeners also experienced silence and pink noise as control conditions. While brainwaves had previously been known to converge with external stimuli with repetition rates in the order of 10-40 Hz, this work found forms of convergence to periodic signals as low as 2 Hz, and throughout the beta/gamma range spanning 13-44 Hz.

Similarly, in a book chapter from Siever & Collura (2017: pp. 51-95) describing clinical usage of *audio-visual entrainment* (AVE), we find that human brainwaves converge to both sound- and light-based stimuli; the authors note that "The frequencies of the lights and tones [used in this context] are in the most common brain wave frequencies, typically ranging from 1 to 40 Hz". In these clinical situations, where brainwaves are stimulated to converge with multiple forms of sensory input, convergence to these treatments has been noted for effects related to dissociation/hypnotic induction, increased cerebral blood-flow, and increased neurotransmitters, among others. Importantly, the authors direct attention to the fact that we encounter forms of audio-visual stimulation constantly in everyday life (e.g., watching TV) where no such effects are observed. They explain that, much like the periodicity described in the study from Will & Berg (2007), it is regularity in these signals that appears to drive convergence. Specifically, Siever & Collura note that "when AVE is randomized at ±1 Hz (for instance, 10 Hz varying randomly between 9 and 11 Hz), entrainment is reported to provide a significant clinical impact, while at ±2 Hz, the clinical effect is poor, and at ±3 Hz, the clinical effect is lost." Therefore, it appears that one factor playing a role in such convergence is the requirement for stimuli to remain relatively consistent and rhythmic.

Another example of physiological convergence is rooted in motor control. Miyake (2009) describes how "Everyone has probably experienced the phenomenon where their footsteps unconsciously synchronize with their partner while walking together: This interpersonal synchronization of body motion has been widely observed and is significant in the context of social psychology." Such motor-synchronization, referred to by Miyake as *mutual entrainment,* is extremely common and exemplifies how humans are capable of converging automatically in physical movement as well; it also reinforces the potential importance of social factors in human convergence.

However, the degree to which people are sensitive to context is still an open question, and through further experimentation we are becoming increasingly aware of what appears to be a rather acute sensitivity to various forms of external information. In fact, an innovative study from Murray-Smith, Ramsay, Garrod, Jackson, & Musizza (2007) explores the possibility of gait alignment during mobile phone conversations. The authors were interested in whether or not talkers' walking patterns would converge despite communicating from different physical spaces (that is, across a park and unseen to each other), and if any patterns observed might be augmented through vibrotactile feedback expressing the footsteps of their conversation partner. Participants were outfitted with a backpack containing a laptop computer to record their speech; a mobile phone; a Bluetooth headphone/microphone combination unit that aided in both capturing speech and conversing through the mobile phone; and a PDA (tablet) used to provide walking-feedback (through vibration), log movement data, and convey experimental instructions. Treatments included script reading, picture description, and free conversation, which could involve either vibrations (via the PDA) reflecting the subject's own gait, or that of their interlocutor. The authors hypothesized that (1) gait

could be influenced through vibrotactile feedback reflecting footsteps of the interlocutor (referred to as *crosstalk*), and (2) progressively more free forms of conversation would also result in relatively increased gait-alignment. Under certain conditions both of these hypotheses were shown to be correct. However, the condition eliciting the strongest gait-synchronization involved free conversation and no crosstalk. This finding suggests that conversation partners are sensitive to low-level auditory information, available through *some* channel in a mobile phone conversation, which reflects the gait of their interlocutor. Moreover, such information is not only transmitted and processed, but is capable of influencing motor activity for the speaker/hearer.

Finally, moving on to the role of convergence in behavioural studies, we also find forms of alignment that appear to influence human behaviour in the social psychology literature. For instance, a Master's thesis from Down (2009) shows that the tempo of background music in a bar is inversely proportionate to patrons' spending habits – that is, relatively slower tempoed music resulted in relatively higher sales at the bar, all other things being equal. In this study musical tempo had no reliable effect on duration of stay. Another study from Roballey, McGreevy, Rongo, Schwantes, Steger, Wininger, & Gardner, (1985) explored the effects of musical tempo on restaurant patrons' eating habits through three test conditions: Music with a fast tempo, a slow tempo, and no music. Their study showed a significant effect of Bites-Per-Minute, which was largest for the fast-tempo condition; and, much like the study from Down (2009), there was no effect of tempo on duration of stay. Work from Smith and Curnow (1966) tested for effects of loudness in a study where music levels spanned loud to soft in large supermarkets. Their results showed that significantly less time was spent in markets during the louder conditions, and that the rate of spending also increased. This may, at first, seem incongruent with the findings of Down and Roballey et al. – however, it should be noted that time in restaurants and bars is typically much more of a social event than is time spent shopping for groceries. As a result, effects may be expected to differ across these situations. As a final example, it appears that even sensory perception can be similarly influenced under some circumstances. One study from Kantono, Hamid, Shepherd, Yoo, Carr, & Grazioli (2016) investigated whether or not self-rated preferences for background music could be used to predict perceived pleasantness of three types of chocolate gelati. Generally, it was found that perceived pleasantness of the dessert increased along with reported musical pleasantness. Thus, patrons' perceptions in gustation appear to have converged with their preferences for auditory input.

In brief interim summary, the research described above has shown that convergence appears to be ubiquitous in nature, and that it need not involve agency. We have seen convergence in inanimate objects, insects, and in both human and non-human animals. Previous work has shown that convergence can happen automatically but can also be influenced by social factors. Moreover, in human subjects, we have seen convergence influencing physiology/biology, neurological patterns, motor control, behaviour, and even sensory perception. Finally, it seems that convergent behaviours are at least sometimes best elicited through patterns that are rhythmic and regular. With this in mind, I will next explore human-convergence as it directly relates to language use.

1.2.4 Linguistic-Specific Convergence

1.2.4.1 Non-Acoustic Linguistic Convergence

Focusing next on work related specifically to language use, speakers have been known to converge in a variety of linguistic ways. For example, Brennan & Clark (1996) describe *lexical entrainment* where speakers alter word-choices to pattern more like those of their speech partners. The authors suggest that this form of convergence is largely the product of speakers needing a way to refer to a single object through a shared conceptualization. Similarly, Garrod and Doherty (1994) describe how an experimentally established community quickly converges on common descriptors for the scene in a maze-game task; in this study, language adopted by larger sub-sects within the group was used more consistently than scheme-descriptors adopted by more isolated pairs of players (again, reinforcing the influence of a relatively stronger sociolinguistic network). Building on works like those just mentioned, van der Wege (2009) provides evidence for a complementary phenomenon she calls *lexical differentiation*. She argues that in the context of lexical entrainment, speakers "contrast any new referents against this previously established set, thereby avoiding applying the same reference phrase to refer to different referents". Therefore, it seems that in spontaneous conversation speakers converge with regard to referents, and then at times *jointly* diverge in form from those referents to maximize contrast in the name of clarity when ascribing new referents as communication continues.

Non-acoustic linguistic convergence is not limited to lexical choices however. Giles, Coupland & Coupland (1991: pp. 41, 49) and Branigan, Pickering, McLean & Cleland (2007) describe speakers' syntactic constructions as also becoming more alike over time. This study from Branigan et al. recounts pairs of speakers taking turns describing pictures to each other; though, one of the speakers was in fact a confederate of the authors, and produced scripted descriptions that varied systematically in syntactic structure. Subsequent syntactic structures from non-confederate participants were found to be influenced by those of the confederate.

Elements of convergence have been attested through morphological experimentation as well. For example, a study from Beckner, Racz, Hay, Brandstetter & Bartneck (2016) lists four specific aims in the context of morphological convergence, all of which are at least tangentially relevant to the present work: (1) They test whether or not human participants converge in verbal morphology (that is, be influenced by a partner's use of regular vs. irregular forms when conjugating non-word verbs); (2) They compare potential for linguistic convergence to conformity on a non-linguistic task (a picture assessment task vs. the verbal/processing task rooted in morphology), thereby exploring the tendencies for each participant to convergence generally; (3) They test for generalized patterns of linguistic convergence as opposed to effects which may be observed for only single items (i.e., application to novel items, and not only similarly producing items already mentioned by their partner); and, finally (4) They explore the social contexts in which conformity does and does not occur – specifically, whether or not there will be any difference in convergence to human partners vs. humanoid robots. The authors describe significant levels of morphological-conformity, though this conformity was restricted to the human-peer condition only. Conformity levels resembled baseline performance when speakers interacted with the robots. It was also found that convergence in the visual (i.e., non-linguistic) task reflected conformity in the morphological task, though this effect also

appeared to be driven by responses from interactions with the human peers. Put simply, participants who were relatively more susceptible to peer-pressure in one condition showed similar susceptibility in the other, but could only be peer-pressured by the humans (and not the robots). And finally, at least in the human-peer condition, participants were found to converge as the experiment continued. Therefore, it seems that convergent-patterns were adopted and generalized to novel items. The fact that morphological convergence was detected in the study, though only in the human-peer condition "supports a view of linguistic convergence as a deeply social process. The level of linguistic conformity displayed by individuals is related to their degree of conformity in nonlinguistic tasks, suggesting that there are individual propensities toward peer imitation that transcend modalities" (Beckner et al., 2016). Put another way, certain individuals appear more likely to converge than others, and it seems that social factors mediate such convergent behaviours – at least to some degree.

1.2.4.2 Acoustic Convergence in Language

When considering the possibility of *acoustic/phonetic* convergence in speech production, studies from both Evans & Iverson (2007) and from Pardo, Gibbons, Suppes & Krauss (2012) explain that the general speech patterns of university students changed to become more like ambient language use over time. Where Evans & Iverson describe students' vowels generally becoming more like the local dialect, Pardo et al. explain that recordings of students were judged to be relatively more alike after the speakers cohabitated during schooling; similar effects are well-attested in the literature, and in this context appeared strongest when flatmates self-reported higher levels of "closeness" (that is, relatively stronger personal relationships). Similarly, Babel (2012) describes more specific acoustic variation, arguing that certain vowels in her study were more likely to be shifted in accommodation (low > high vowels) and, not unlike the findings of Pardo et al., explains that one is more likely to accommodate when they hold positive opinions of their speech partner (in this study gauged via attractiveness ratings).

Next, focusing more directly on fine phonetic detail within speech, Pardo (2013) reviews the current literature investigating acoustic-phonetic convergence between speakers and cites a score of research implicating duration, speaking rate, F0 or intonation contour, intensity, voice quality, vowel spectra, voice onset time (VOT), lip aperture, and individual phonemic variants as known areas of convergence (see Pardo, 2013 for references). Because the scope of the present work is limited to investigation of acoustic-phonetic convergence to ambient noise, areas of investigation, and therefore, review, will be restricted to studies exploring convergence in duration/speech rate, F0, and intensity. This decision was made because the remaining acoustic attributes (i.e., voice quality, vowel spectra, VOT, lip aperture, and phonemic variation) cannot be directly accessed or compared through most non-speech noise.

Among other acoustic dimensions, one study from Levitan & Hirschberg (2011) tests for convergence to voice-pitch, intensity, and speaking rate. This study makes use of the Columbia Games Corpus (comprised of twelve spontaneous dyadic conversations) to investigate potential acoustic-phonetic convergence at both the session-level and at the turn-level. The study focuses primarily on how localized this form of convergence may be: specifically, the authors ask whether or not acoustic convergence is an ongoing process of coordination that may improve over the course of a conversation; at what point speakers adapt; and if speech patterns become more alike in absolute or

relative terms. To this end, the authors distinguish between two strategies for speech characteristics becoming more alike: (1) *Proximity* refers to similarity resulting from a single coordinating step at a dialogue's onset, and (2) *Convergence*, which is operationally defined as a process whereby similarity increases over the course of a conversation (i.e., an ongoing process). Results indicate that speakers align in these three acoustic dimensions at both the session- and the turn-level, and describe entrainment as "a dynamic process of continuous matching at turn exchanges, even in dimensions that do not display session-level proximity." It therefore appears that convergence in speech production can be observed at both the micro and macro levels.

Looking at these acoustic dimensions separately, another study from Webb (1972: pp. 115-133) investigates convergence in speech rate (measured as syllables per minute) between an interviewer and an interviewee. The interviewer in this study was, in fact, automated – that is to say, a collection of pre-recorded statements and questions. These statements and questions were played to the interviewee from an adjacent room, where both the separation and pre-recorded questions served to control for any unintended influence of interviewer gesture and the potential for an interviewee to influence interviewer speech rate, respectively. Four test conditions affecting interviewer rate included (1) High rate with short pauses, (2) High rate with long pauses, (3) Low rate with short pauses, and (4) Low rate with long pauses. Webb found that participant speech rates in conditions 1+2 were significantly faster than in 3+4, and that the differences between 1+2 and between 3+4 were non-significant. Thus, while effects were generally found for speech rate, Webb concluded that relatively increasing or decreasing pause durations did not influence subjects' speech rate in any reliable way.

Work from Natale (1975) explores the potential for convergence to speech-intensity; how any such convergence may be influenced by social desirability; and, if tendencies toward convergence between speakers increase over time (cf. Levitan & Hirschberg, 2011: *proximity* vs. *convergence*). The study explores these questions through two experiments. In the first, once again taking on an interview format, subjects heard the interviewer through loudspeakers at one of three loudness ranges: (1) 80-83 dB SPL, (2) 86-89 dB SPL, and (3) 92-95 dB SPL. Unlike the technique used by Webb (1972), the interviewer in this study produced speech on-site, in real time, where intensity levels of the interviewer were restricted to a 4 dB range through use of technology. The interviewer and subjects were in separate sound booths, out of each other's view, in order to control for any unintended influence via gesture, posture, etc.. The second experiment involved unstructured conversation with each conversation pair taking part in three 1-hour conversations over different sessions. In this study, both talkers were in the same room and no headphones or loudspeakers were used, though each participant was outfitted with a unidirectional, lavalier microphone. There were no intensity-restrictions imposed in the second experiment and participants were therefore required to respond to a full range of speech behaviour in this context. Prior to the first session in the second experiment each participant completed the Marlowe-Crowne social desirability scale, which is meant to quantify one's desire for social approval. Again, participants were out of each other's view (here, separated by a curtain) in order to avoid any unintended influence.

Results from experiment 1 show that increasing or decreasing intensity levels of the interviewer generally resulted in corresponding changes in participants' vocal intensity. Results from the second study reaffirm convergence to speaker intensity – however, an interaction between participants' Marlowe-Crowne scores and

observed changes in intensity suggest that speakers with higher social desirability were contributing relatively more to the convergence observed in each dyad than were speakers with lower scores on the Marlowe-Crowne scale (thus, reinforcing the importance of both social interaction and hierarchy). Moreover, increasing levels of convergence were found when exploring mean intensity levels across the three sessions; however, I would interpret this latter finding with some caution, as this portion of the analysis was inherently confounded (though, in the end it is likely of little consequence). When extracting intensity levels from each session, the experimenter took mean intensity from the first ten minutes of session 1, the middle ten minutes from session 2, and the final ten minutes from session three. It could be the case that convergence is increasing at a much quicker rate (i.e., within a single session) than on the macro level (that is, across sessions), but it is surely the case that convergence was increasing on *at least one* of these temporal levels. Therefore, it appears that increased social contact contributes to increased alignment, at least in some contexts.

When considering potential for convergence to voice-pitch patterns, a number of studies have shown this to be an area where speakers align (e.g., Lieberman, 1967; Gregory, 1983; Gregory, 1990). One study from Gregory, Dagan, & Webster (1997) investigates accommodation to voice-pitch through a production component, and further assesses how alignment of F0 may influence social interaction through a perceptual component of the study. This work also utilizes an interview/free-conversation paradigm when eliciting speech from participants. However, regarding measurement of interaction *quality*, this peripheral role of voice pitch was tested by strategically filtering out F0-information (high-pass at 550 Hz) from one of the two speakers in 1/3 of the dyadic pairs tested. The study further included a low-pass condition as a third treatment, to which 1/3 of the dyads were also assigned (low pass at 1000 Hz). The authors hypothesized that removal of F0-information would result in a relative decrease in voice-pitch accommodation. They also hypothesized that removal of this information would impact the quality of the conversation for those pairs in a way that would be qualitatively detectable by judges hearing *unfiltered* versions of those recordings after the fact. Judgments for the perceptual component were restricted to binary positive/negative responses on a semantic differential scale (e.g., "good/bad, pleasant/unpleasant, and friendly/unfriendly").

Analysis of the production portion of the study confirms the authors' first hypothesis, and provides evidence for an important communicative role of F0 in spoken language. Both the unaltered and low-pass dyads were found to exhibit relative convergence in voice-pitch, whereas the high-pass dyads with voice-pitch information removed showed no such alignment. Next, discussing how conversation variants were perceived by others, the semantic differential ratings supplied were condensed into three main factors. Firstly, with regard to *Factor1 - Social* (e.g., friendly/unfriendly, pleasant/unpleasant) judgments: These loadings for the unaltered group were highly positive, whereas they were all negative for the high-pass group with F0 removed. Though not as strong as for the unaltered group, the authors describe a general correspondence between judgments of the low-pass and the unaltered conversations, where judgments were also mostly positive. No significant differences were observed for judgments regarding *Factor2 - Substantive* (e.g., valuable/worthless, relevant/irrelevant). The authors explain this absence of differences is rooted in semantics: put simply, this factor focuses primarily on verbal content, which was not disrupted in any meaningful way through the treatments. Finally, judgments for *Factor3 - Potency* (e.g., aggressive/timid, dominant/submissive) were much like with Factor1; however, in this case the low-passed

conversations were judged to be more like the high-pass conversations. In short, the unaltered conversations were judged to be highly positive, and both filtered conditions were judged more negatively. The authors explain that this factor is focused primarily on power ratings, and that positive information regarding power between talkers cannot be conveyed effectively when F0-information has been removed. Moreover, based on these results, the authors speculate that such information may not be conveyed effectively when any portion of the voice at any bandwidth has been altered. When considered together however, these results support F0 as contributing to the social/evaluative quality of interactions, and imparting information regarding the quality of the communications context. It is also important to note that, the specifics of judgments aside, there was in fact a qualitative difference that speakers were able to detect from these conversations with varying degrees of information removed at the time of recording, even when judges experienced all speech in an unaltered form. Much like we saw in the study from Murray-Smith et al. (2007) (re: synchronization of gait during mobile phone conversations), these findings reinforce the notion of human listeners as extremely sensitive to various channels of (not necessarily linguistic) information available through spoken language.

Speaker and listener gender has also been shown to play an important role in shaping linguistic convergence. When describing general gender-based trends, Namy, Nygaard & Sauerteig (2002) explain that "In general, women are more likely to accommodate to a conversational partner than are men, and both male and female participants tend to accommodate more to male than to female partners." Gender-based trends, however, are often not so simple. For example, Bilous & Krauss (1988) explore convergence in same- and mixed-gender dyads during semi-structured conversation, looking at convergence in speech to various aspects of conversations (e.g., interruptions, pauses, laughter, listener back-channel responses, etc.). The authors found that male and female speakers converged in some aspects of conversation, diverged on others, while showing no effect on others still – and these results were complicated further as a function of dyad type (that is, single-gender vs. mixed-gender). Bilous & Krauss explain that on nearly all of the measures assessed, both male and female conversation-behaviour was more alike in a mixed-gender context than in single-gender conversation, though observed behaviours were extremely complex and could not be reduced to a simple generalization. When describing their findings, the authors state "while it is clear that males' and females' conversational behaviour varies significantly with the gender of their conversational partners, the form that accommodation takes [i.e., convergence vs. divergence] depends upon the particular speech index that is being considered." Interestingly, at an extreme in mixed-gender dyads, both male and female speakers at times exhibited what the authors describe as a *hyper-convergence*, where speakers' production patterns moved so far in the direction of the gender-norms of their conversation partner that a reversal of the original pattern was observed. One (speculative) means of explanation offered for the trends observed in this work involves the drive(s) for convergent and divergent behaviours; the authors suggest these changes may in part be driven by level of involvement, a desire to dominate the conversation, and impression-management while working with previously unacquainted peers. By this account, convergence in speech is largely shaped by a social component; though, is also clearly automatic to some degree, at least insofar as much of these behaviours are not consciously engaged by the speaker.

Other work from Willemyns, Gallois, Callan & Pittam (1997) investigates accent-accommodation in a job-interview context, focusing on interviewer accent and gender. In this study interviewers were either female or male, and spoke with either a *cultivated* Australian accent or a *broad* Australian accent (confederates were trained to speak with comparable degrees of each accent). Interesting gender-based effects were found with regard to both speech production and perception: First, even though the amount of accent had been matched across speakers, male interviewers were generally perceived by interviewees to have slightly more broad accents than their female counterparts. The authors describe this perceptual effect as likely driven by distortions based in gender-role stereotypes and expectations. Simply, subjects *expected* the male interviewers to have relatively stronger accents, and so were primed to recognize accents as relatively stronger. Second, with regard to production, the female speakers were found to be more accommodative to both accents, whereas the male speakers were often found to diverge from the cultivated accent. Willemyns et al. argue that this divergence is likely the product of male speakers identifying with the broad accent more than the cultivated, and thus taking on an out-group accent to demonstrate non-identification with the interviewer. Another point of note also involves the broad-accent condition, in which participants generally spoke with broader accents than in other conditions – though, speakers' self-ratings indicated they were unaware that they had adopted this relatively broader accent in the context. This study provides further evidence demonstrating that convergence in speech is at least partially automatic, though clearly is shaped to some extent by a social motives and interaction as well.

Beyond gender-based effects, language attitudes have also been shown to influence convergent behaviours (e.g., Love & Walker, 2012; Giles, Coupland & Coupland, 1991; Drager & Kirtley, 2016: pp. 9). For example, Drager, Hay, and Walker (2010) describe a production-based study exploring the degree to which speakers of New Zealand English (NZE) may shift their productions to be more Australian-like when exposed to polarized facts about Australia. It was hypothesized that speaker's vowel realizations would be influenced differently by 'good' vs. 'bad' facts about Australia, and that convergent and divergent behaviours would be shaped to some degree by speakers' pre-existing biases toward Australia rooted in sport. Productions were tested through word lists focusing on the vowels well known to distinguish NZE from Australian English. These lists were read aloud before speakers read one of the fact-lists raising a polarized concept of Australia; speakers would then read the word list once again, which could be compared in a straightforward way to their pre-fact-list productions. Finally, the fact-list would be read aloud once more. Following test conditions, all participants filled out a questionnaire that included questions about sporting interests for later analyses. It is remarkable that in this study no Australian speech is presented for speakers to converge/diverge with – the study tests whether or not evoking *the concept* of Australia is sufficient to elicit convergent/divergent behaviours. Results indicate that the concept alone was enough to elicit altered production patterns. Moreover, as predicted, convergent and divergent behaviours appear to have been shaped by attitudes about Australia that had been coloured by the polarized fact-lists and sporting rivalries. For non-sports-fans, positive facts about Australia appeared to evoke positive attitudes, and subsequently more Australian-like vowel production, whereas the negative facts resulted in more NZE-like vowel productions. The sports fans, however, exhibited different effects: The 'good' facts about Australia resulted in more Kiwi-like vowel productions; the 'bad' facts, however, resulted in less out-group productions (or more convergent behaviour), which the authors explain as

the product of a shifted baseline due to speakers' attitudes. Due to the sporting rivalry, is it believed the default attitude toward Australia is negative for the sports fans, who become increasingly defensive (or perhaps nationalistic) when the 'good' facts are introduced, thus evoking more emphatic or exaggerated NZE vowel productions. When explaining these effects the authors' comments are much in line with Babel (2009), where they state "…that speech convergence and divergence can occur as a result of one's active construction of identity as well as cognitive processes which are more automatic." Thus, this study provides further evidence that while certain aspects of convergence and divergence may be automatic, there can be a powerful influence of experience and social motivations.

Moving on to the only known works exploring acoustic-phonetic convergence to *ambient* speech (recall ambient speech has been operationally defined above as "…spoken language experienced by the talker where there is no explicit interaction with an interlocutor, nor is the background speech related in any direct way to the experimental task"), very few known studies exist. Sancier & Fowler (1997) investigate variation in Voice Onset Time (VOT) through recordings of a speaker alternating between extended periods of time in two countries where different languages are spoken (i.e., American English and Brazilian Portuguese). While the paper does investigate and find changes to speech patterns which appear to be driven by ambient language exposure, such changes can be categorized as more phonemic in nature, where phonetic targets appear to be altered – that is, speech patterns are modified to more closely resemble those of the new speech community after moving country, likely in the name of more effective message transmission (e.g., Adank, Hagoort, & Bekkering, 2010), or perhaps economy of effort. It seems a different process in this context than entraining to the envelope of some acoustic variable over time, where here different exemplars (e.g., Pierrehumbert, 2001) are re-weighted for the speaker given increased/decreased exposure to an alternate language, and the specific phonetic trends are gradually altered accordingly. On this subject Babel (2009) argues that "dialect change is essentially long-term accommodation to new speech patterns". The present work investigates only more immediate, short term forms of convergence – though, I will return to Sancier & Fowler (2007) later on when outlining various mechanisms that lead to convergence in speech in section 1.2.5.

The best-known (and only?) study investigating acoustic-phonetic convergence to ambient speech comes from Delvaux & Soquet (2007a, 2007b). These two papers describe the same work, one as an expanded journal article (2007a) and the other as a truncated proceedings paper for the International Congress of Phonetic Sciences (2007b). I will therefore only discuss the full article (2007a) in this review. Two related experiments were carried out: Experiment 1 involved a naming task where French native speakers from two regiolects within Belgium (n = 2 x Liège and n = 2 x Brussels) experienced two conditions. In the first, referred to as "Record", participants were asked to read stimuli aloud that were presented on a computer monitor; these productions served both as a baseline for later analyses, and as stimuli for other participants in the second condition. In the second condition (referred to as "Test"), participants were exposed via loudspeakers to productions from the regiolect which was not their own. The second condition took place approximately two weeks after the first. In both sessions participants sat at a table directly across from a computer screen with loudspeakers at each side. Reading materials selected for dialectally telling phonemic variants were presented on the computer monitor and participants were directed to read them aloud in a carrier sentence when their turn came round. Specifically, in the second session an arrow would appear on the

computer monitor either pointing to the participant, or to one of the speakers at the left or right side of the table. If the arrow pointed to a loudspeaker then the participant would hear a pre-recorded production of the stimulus through that speaker (exemplifying the other regiolect, as captured in the first session). If the arrow pointed to the participant, they were then expected to produce the stimulus within a carrier sentence. It was theorized that exposure to the non-native regiolect would elicit convergence in vowel productions, where speech would begin to better resemble the other regiolect as the experiment progressed. Comparing productions from the Record condition to those captured during the Test condition, the experimenters found that Test-productions of the variants of interest (/o, i, s/) had changed phonetically in ways that reduced the phonetic distance between the regiolects. Thus, it appears that participants' production patterns were influenced by ambient speech in real time, to change in ways that made them better resemble that ambient speech.

Experiment 2 served largely as a replication of the first experiment, though testing for effects with Mons speakers (n = 8) hearing a Liège speaker. This replication, however, also introduced a third, post-test condition (a repeated Record condition) in order to explore whether or not convergent effects would persist after the conclusion of the Test condition. Once again, when comparing productions from the initial Record condition to those from the Test condition, participants' speech changed in ways that became quantifiably more like those of the ambient regiolect. The effects observed in the Test condition subsided during the third condition – that is, the second Record condition – though not completely. Thus, experiment 2 shows the influence of the ambient stimuli appeared to persist, at least somewhat, even in the absence of those stimuli (cf. Hay et al., 2017).

In brief summary, we have seen that language users converge in many ways outside of the realm of acoustics, including lexical choices, syntactic structures, and even in use of morphology under certain conditions. We have also seen that human speakers at times converge with conversation partners in many acoustic dimensions; however, because the focus of the present work is restricted to potential convergence to ambient noise, I have limited discussion for the most part to convergence observed in voice-pitch, speech rate, and intensity, all of which are well attested in the literature. We have seen that sometimes these forms of acoustic-phonetic convergence are influenced by social motivations, as in the studies from Natale (1975) and Gregory, Dagan, & Webster (1997). I have also discussed the only known work exploring convergence to ambient speech, which provides evidence that speakers' productions can be influenced by ambient talkers, and that such effects can persist for some time after the test conditions have ended. At this point, it makes sense to next discuss the different mechanisms argued to drive convergence. This discussion is available directly below in section 1.2.5.

1.2.5 Different Mechanisms Driving Linguistic Convergence

When considering speech production in the context of convergence, it is important to recognize that linguistic entrainment can be divided into what is often described as *social* and *automatic* forms. I use the term 'social convergence' when referring to the type of effects observed by Eckert (2008), Drager (2011) Giles (1974), and Marx (2002) where, for example, sociolinguistic variables can be used to signal shared identity, or to signal speakers as different from some other group. Conversely, I use the term 'automatic convergence' to describe necessarily more

reflexive forms of alignment, the effects of which are often (at least consciously) unnoticed by speakers and listeners, and are relatively more automatic in nature (e.g., Lakin, Jefferies, Cheng & Chartrand, 2003). The present work focuses primarily on automatic forms of convergence.

I further divide such automatic convergence into three main sub-groups, differing by both the function of the convergence and the duration for which an effect may persist. The first type involves the influence of a dominant ambient-language and is not unlike environmental influences described in phonological acquisition (e.g., Chambers, 2002; Warren, 2001: pp. 84), second language acquisition (e.g., Derwing & Munro, 2013; Klein, 1995 – cited within Bongaerts, 1999: pp. 154), and some work exploring degrees of native-likeness in speech (e.g., Gnevsheva, 2017). In this form of convergence speakers' longer-term phonetic trends and phonological categories are shaped over time to become more like those of the speakers around them (see Pierrehumbert (e.g., 2001) for discussion on Exemplar Theory). While phonological patterning can be altered consciously in the name of more effective message transmission or identity expression, acoustic/phonetic changes and the re-shaping of phonological categories are often the unavoidable and automatic byproduct of prolonged exposure to, or interaction with a speech community (e.g., Sancier & Fowler, 1997). In other words, following sufficient experience, speakers' productions become relatively more like those of the speech community – and productions often change in this way to some extent whether the speaker was consciously aiming to reshape them or not. However, effects explored within the present work appear to influence speech in real-time; such change is therefore unlike the relatively longer-term reshaping of phonemic categories through modified exemplars/prototypes, a process which requires more time and exposure than participants are currently receiving in this series of experiments.

Another form of automatic convergence can be found through a paradigm known as Shadowing (e.g., Goldinger, 1998), and Mimicry (e.g., Flege & Hammond, 1982). These effects are not known to reshape mental representations of phonological categories, but instead involve speech production patterns changing with more immediacy. Shadowing describes a situation where speakers are asked to repeat speech that they hear as quickly as possible; this process often results in not only repeated words and phrases, but speakers also reproducing some of the prosodic information from the original speech as well (hence the term '*Mimicry'* sometimes used in this context). Related effects have been shown to persist in both word duration and vowel quality, where a speaker's productions of the specific words and phrases are more like those *shadowed* for some time following the initial session (Goldinger & Azuma, 2004). However, these effects diminish over time and have not been shown to persist for as long as those described in the studies on ambient language (cf. Exemplar Theory). Simply, the length of time that these effects are observed in later productions appears to be far shorter than those associated with renovated phonemic representations. They do, however, appear to persist for more time than the immediate changes to speech observed and described in the experiments designed for the present work.

Finally, automatic convergence can be realized as changes in real-time, where acoustic attributes of speech vary along with some comparable stream of information within a given stimulus. In this case, changes to acoustic characteristics like voice-pitch or speaker intensity, for example, ebb and flow with those of the background noise (e.g., music). While it is currently unknown exactly how long such convergent effects may persist (cf. Delvaux & Soquet, 2007a), the experiments described within this dissertation indicate that these types of changes appear to be

both immediate (that is, changing along with the stimulus) and fragile (in that they last only until the speaker experiences something new to differently affect their speech, or possibly until the stimulus is no longer present). As mentioned above, outside of tangentially related work on the Lombard and Fletcher effects (described in section 1.2.6) only two related studies from Delvaux and Soquet (2007a, 2007b) have been published on this type of convergence, both of which focus exclusively on background *speech* as noise.

However, please recall that Babel (2009) argues that neither a completely automatic nor a solely social theory of accommodation was suitable to explain her data. In a statement that sits well with the literature described above, she states:

> "The selective nature of imitation from both a phonetic and social perspective is evidence that purely automatic theories of imitation are wrong. Imitation is mediated to some extent by both social and linguistic factors. However, participants were unaware of their imitative behaviours and the fact that imitation targeted specific vowels suggests that spontaneous phonetic imitation is well beyond a talker's conscious control. This fact also shows that a social theory in the vein of sociolinguistics or [Communication Accommodation Theory] cannot be completely right."

Thus, it appears that convergence in speech production may not be wholly driven by automatic nor social factors. Instead, this process may better be thought of as something more continuous, which varies by situation in the degree to which it is driven by reflexive vs. social mechanisms.

Because there is no previous work directly testing for acoustic-phonetic convergence to ambient noise, I will next briefly review literature on the Lombard effect (Lombard, 1911) and some Speech-in-Noise research. The Lombard effect is often cited in studies exploring convergence to intensity and, at least in some respects, can be considered a form of acoustic entrainment. Moreover, because convergence to background noise *must* involve some form of speech production with background noise that might be entrained to, both of these literatures served to guide decision-making while designing experiments for the present work. This brief review is available below.

1.2.6 The Lombard Effect, and Some Other Work on Speech-in-Noise

A large body of work concerning speech-in-noise also proves informative while exploring the potential for acoustic entrainment, with the earliest relevant claim coming from Étienne Lombard (1911). Lombard found that speakers' vocal intensity increases when competing with environmental noise of a sufficient level. This change in speaker loudness eventually came to be known as the Lombard effect. Subsequent works, however, have shown that acoustic characteristics beyond vocal loudness often also change under Lombard conditions. The effect has since been associated with increased fundamental frequency, increased word and segment duration, relatively higher F1 and lower F2 values, and a leveling of spectral tilt (e.g., Cooke & Lu, 2010; Draegert, 1951; Brumm & Zollinger, 2011;

Hay, et al., 2017). Cooke and Lu (2010) and Stowe and Golub (2013) also explain that acoustic variation in Lombard speech can also differ as a function of noise type. That is to say, inequivalent acoustic changes, beyond intensity variation, have been observed in Lombard speech when conditions involve one type of noise mask instead of another.

Thus, it has been demonstrated clearly that speakers' productions reflect changes encountered in the intensity envelope of a speech-masking noise, at least to some extent. But the fact that other acoustic properties – not *necessarily* related to intensity – also often change in production under noisy conditions suggests that some of these changes may be driven by other specific acoustic information within that noise. For example, rising F1 and lowering F2 levels are observed only intermittently in the Lombard literature, where this variation could perhaps be somewhat specific to the noise masks selected as stimuli for those studies. Maybe some secondary changes (beyond intensity variation) are in fact the product of speaker-entrainment to (or divergence from) some component(s) within the masking noise? While this hypothesis is not unreasonable, the extent to which acoustic changes in production, such as rising F1, can be explained by entrainment as opposed to effects driven by speaker physiology, is at present unknown. There is some work though, exploring the acoustic and physiological correlates of speaker-loudness through voluntary/instructed degrees of vocal loudness as compared to more involuntary changes driven by background noise and the Lombard effect. These studies indicate that louder speech is often characterized by increased jaw openings, and thus a relatively lower tongue height, resulting in the relatively higher first formant values often observed (Shulman, 1989; Huber, Stathopoulos, Curione, Ash & Johnson, 1999; Tam, 2017). Such effects may be exaggerated when increased vocal loudness is achieved through more agentive efforts.

In much the same way that the Lombard effect varies in its specifics by noise type, we have also seen related acoustic changes that call into question the roles of experience and expectation. For instance, Amazi & Garber (1982) describe an experiment where young children and adults experienced comparable Lombard-inducing conditions while completing labeling and storytelling tasks. The authors explain that all subjects increased vocal intensity in the presence of noise. However, where the level of increase varied for adult speakers by task, presumably due to the communicative element inherent to a storytelling task, the children were relatively uniform in their increase regardless of task. Also related to previous experience, other work from Hay et al. (2017) explores the influence of physical environment on speech production and processing. Participants experienced different noise conditions i.e., no noise vs. restaurant noise vs. car noise (simulating the noise of an automobile running on the highway) in different physical settings; that is, each auditory stimulus was encountered in either a laboratory setting or within a parked car. As might be expected, the car noise elicited Lombard-like speech in both physical settings when presented at an appropriate level. However, where in the laboratory-setting speakers showed perceivable acoustic changes when comparing the quiet and car-noise conditions, within the car the same quiet condition elicited speech much like that produced in car noise. It appears that one underlying mechanism for variance in both studies is *previous experience* – it is possible that speakers learn to expect road and engine noise in a car, and as a result of this expectation preemptively compensate for the expected ambient noise. Similarly, perhaps the young children described by Amazi and Garber have yet to learn some distinction, or make some cognitive connection in the way of inhibition, which alters the use of resources in auditory processing as a function of other cognitive demands – most

likely, in order to increase speech intelligibility. When considering results from these two studies together we might draw two important inferences: 1) That the acoustic changes brought on by certain auditory conditions *may be learned or malleable* to some extent, and 2) That such changes may not be elicited *when one lacks the appropriate experience and/or expectation* to reshape production (e.g., without sufficient experience hearing the appropriate noise in the car-setting, speakers would have no previous experience with that noise to draw from that might drive the pre-emptive compensations). If these inferences hold, then participants' experience must be taken into account in phonetic studies: for in some contexts it could potentially explain the presence or absence of specific changes to fine phonetic detail (and, therefore, predictable variation) in speech production.

Psycholinguistic literature focussed on *speech shadowing* also describes ways in which experience and expectation might influence the acoustics of spoken language. Goldinger (1998) defines shadowing as a task where participants hear spoken words and quickly repeat them – these tasks are frequently used to explore theories of lexical access through reaction times, though some researchers have bent this methodology to focus more specifically on acoustics as well. Two very important results were reported by Ryalls & Pisoni (1997) from their study pairing speech-in-noise with a shadowing paradigm: They found that (1) Young children who were relatively older had an advantage over younger speakers when processing speech from multiple talkers presented separately in noise (as opposed to processing speech from only a single talker – hence, a learned generalizability), and (2) That in a shadowing-in-noise task where both young children and adults were relatively slower to repeat words from multiple/variable speakers (again, as compared to repeating speech from a single speaker), *only the children matched durational properties of the speakers' productions*. This can be taken as an example of children entraining to specific acoustic information where adult speakers showed no such effect.

However, in an unpublished shadowing-study mentioned in Goldinger (1998), adult participants were found to match both duration and F0. Given the conflicting results and variable methodologies across studies it is unclear whether young children may be more or less susceptible than adults to some forms of entrainment – but when considered together, it seems that children may not entrain to speech in a shadowing task in precisely the same ways as adult speakers, and it is possible that convergence in this way may be learned/inhibited as appropriate under certain circumstance(s). Note, though, that Pardo (2013) argues acoustic-phonetic convergence is, if anything, inconsistent. In a review of acoustic entrainment she explains that talkers "can converge on one dimension at the same time that they diverge or produce random variation in other dimensions… [for example,] convergence on F0 on one item does not imply the talker will always or only converge on F0". So while it is possible that learned behaviours may account for these conflicting results, perhaps some other contextual factor (such as inter-participant differences, inconsistent behaviours, or even communicative/social aspects specific to the task) may better explain some of them.

It seems worthwhile then, at this point, to briefly discuss and compare background noise as it is traditionally treated in speech-and-noise studies, and how music has been integrated for use as background noise within the present work. Where studies on speech-in-noise typically treat noise as *something to mask speech,* in sociolinguistic research focussing on accommodation the noise (namely other speakers' productions) is most often treated as *a potential point of reference for relative changes* that might be observed in a subject's speech. In other words, the

speech-in-noise paradigm is most often used to explore either 1) how effectively different forms of noise mask human speech, or 2) which specific characteristics of speech might be altered in the name of more effective message transmission – both of which questions are largely listener-driven in their focus. However, in the present work this paradigm is repurposed to use background music as a baseline from which relative changes might be observed in speech productions by comparison. Importantly, the drive for any potential convergence in this context would likely not be listener-driven (i.e., in the name of increased intelligibility) but instead could potentially exist as a means to signal opinions and/or social affiliations, as has often been observed in other accommodation-focussed sociolinguistic studies (e.g., Willemyns et al., 1997; Babel, 2010).

Finally, from a theoretical perspective it seems worthwhile to also consider exactly why certain effects observed through speech-in-noise might occur – that is, are these changes to speech patterns in this context typically listener-driven (e.g., in the name of improved intelligibility) or speaker-driven (e.g., perhaps to reduce speaker effort in production or demands in cognitive processing)? Work on the Lombard effect provides the most clear example of speech-in-noise research which resembles acoustic-phonetic accommodation: specficially, speech productions are well-known to converge with (that is, become *more* like) or diverge from (or become *less* like) the noises a speaker encounters in certain ways. For example, as mentioned above, this effect was first recognized as a form of pattern matching where speakers' vocal intensity was found to get louder and quieter in accordance with competiting noise (Lombard, 1911). Much like the acoustic-phonetic characteristics of two speech partners whose productions become more alike through contact, characteristics of background noise (namely intensity levels) are normally matched by speakers who encounter that noise. In other words, the Lombard response looks very much like a form of convergence insofar as speech begins to reflect properties of the background noise in which it was produced (cf. related forms of acoustic-phonetic convergence described in Pardo, 2013). Similarly, the speech-in-noise literature often describes forms of what might be considered acoustic divergence (again, most often associated with a Lombard response) through spectral changes related to differing noise types. It is known that Lombard responses can vary by noise-type (e.g., Rao & Letowski, 2006; Cooke & Lu, 2010; Stowe & Golub, 2013), where such changes may be interpreted as a form of *divergence* when considered along with economy of effort. For example, these works explain how the spectral characteristics of speech productions often change in ways that might improve intelligibility when competing with certain noise types and not others (e.g., a leveling of spectral tilt when competing against pink vs. white noise, where representation in higher frequency bands would more effectively aid intelligibility when competing against the former). Thus, it seems that speech-in-noise works have shown speakers may, in a sense, diverge from background noise in ways that could support intelligibility as a motivation.

Not unlike work on the Lombard effect, research on *clear speech* suggests intelligibility is often the case. For example, an article from van Summers, Pisoni, Bernacki, Pedlow, & Stokes (1988) describes paired production and perception studies which explore altered production as a function of noise level, and then differences in intelligibility in light of that altered speech. As is now well attested in the clear speech literature, speech produced in noise was associated with increased vocal intensity, duration, voice-pitch, as well as reliable differences in the formant frequencies and short-term spectra of vowels. When speech recordings produced in quiet and in noise were presented at identical signal-to-noise ratios, utterances produced in noise were consistently found to be more

intelligible than those produced in quiet. Similarly, a pilot study described in Hay et al. (2017) explains that when signals produced in quiet and in noise were both presented to listeners in that same noise, these listeners were able to identify whether or not each item was produced in that noise at levels considerably better than chance – even when differences in signal intensity across conditions had been controlled for. If speech is altered in recognizable, acoustic-phonetic ways that improve intelligibility, then it seems possible such efforts may be listener-driven in nature; however, other work suggests such changes may more so be a balance of listener- and speaker-based benefit. Podlubny et al. (2018), for example, describe a series of perception-based experiments where acoustic information is restricted systematically (i.e., exploring isolated contributions of spectral shape, f0 contour, target duration, and time varying intensity) in both absolute manipulation and at gradient SNRs. Crucially, these experiments also explore naturally occurring further signal degradation through grades of phonetic reduction which vary by item. This work shows general improvements to intelligibility as SNRs become more favourable, and that such changes occur despite relatively higher levels of phonetic reduction. While these stimuli were all originally produced in quiet prior to their presentation in noise, it seems possible that frameworks such as Lindblom's Hyper/Hypo theory (e.g., 1990) and Aylett & Turk's Smooth Signal Redundancy Hypothesis (e.g., 2004; 2006) – which describe a fine balance between ease of production for the speaker and contextual expectations based on the listener's knowledge and abilities – may well apply more generally to speech in a variety of adverse listening conditions.

However, there is further crossover between these fields of study – being work rooted in convergence and work studying effects related to speech-in-noise – which might be considered before moving forward, specifically keeping in mind that speech-in-noise work on the Lombard effect is often implicitly or explicitly associated with phonetic convergence, where these two streams of research can be discussed within a single discourse (e.g., Pardo, 2013; Brumm & Zollinger, 2011; Pardo, 2006; Natale, 1975). For example, one study from Tweedy & Culling (2014) explores whether or not different SNR levels might influence the degree to which a speaker might experience a Lombard response. In this study participants experienced simulated conversations (referred to as "live interaction") with an experimentor, where recorded phrases related to a specified topic were played at three SNRs (4, 6, & 8 dB). These conditions were further modulated by three presentation levels for the background noise (65, 62, & 59 dB(A)). The experimentors found that while the level of background noise reliably influenced participants' Lombard response, or *levels of convergence* to signal intensity, the SNRs showed no such effect in this context. It, therefore, appears that a Lombard response (or, convergence to vocal intensity) is influenced selectively by the level of masking noise, but *not* by the relative vocal loudness of the interlocutor competing with that noise. In light of the above-mentioned pilot described by Hay et al. (2017), these results might considered with caution: if subjects could recognize that the speech of their supposed interlocutor was in fact not produced in the same ambient noise in which it is encountered, then it follows that such speech may be processed differently cognitively, and thus related effects may differ in significant ways.

Conversely, Lau (2008) describes a study where pairs of subjects took turns in roles of speaker/listener, where word lists were read by one to the other. The experiment involved white noise played through headphones at 70 dB in four noise-based conditions to mask the speech: (1) neither person in noise, (2) both in noise, (3) speaker only in noise, and (4) listener only in noise. It was hypothesized that speakers may exhibit a Lombard response both

when hearing noise and when they believed the listener was hearing noise. When compared to the no-noise condition, results showed that speakers' mean amplitude was 7.4 dB louder in the *both in noise* condition; 6.3 dB louder in the *speaker only in noise* condition; and, remarkably, 2.16 dB louder in the *listener only in noise* condition. These results fit well with both Lindblom's H&H theory and the Smooth Signal Redundancy Hypothesis, where speaker efforts are argued to be influenced by a speaker's assessment of the listener's knowledge and abilities. If one can assume that a speaker's previous experience in noise may result in a sort of productive empathy, or a preemptive compensation in the name of intelligibility based on knowledge of the listening conditions, then what we might refer to as a sympathetic Lombard response being observed here would not be completely unexpected (cf. Hay et al., 2017).

However, while various forms of convergence to speech and what might be called convergence to signal-intensity (through work on the Lombard effect) are well documented, no work has yet explicitly tested for any forms of convergence to acoustic characteristics of ambient noise. It is therefore currently unknown whether or not speakers will converge to other acoustic properties of ambient noise and, returning to the question of speaker vs. listener-driven changes regarding speech-in-noise, what influence there may or may not be of the non-communicative context. Sociolinguistic work on convergence suggests that increased levels of similarity often occur to win approval (Bell, 1984), or are at times the product of positive opinions of your interlocutor (Babel, 2012); though, it is unknown at present how any such convergence-based effects may be manifest in a context which is inherently non-communicative.

1.2.7 Differences in Performance Related to Experience

While work mentioned above from Hay et al., (2017) shows that acoustic differences in speech production can be driven by personal experience, it might be expected that musicians specifically would respond differently than non-musicians to experimental treatments, given the musical nature of stimuli used throughout this dissertation. In fact, a variety of studies indicate that musicians may perform differently than those who lack the same musical experience in a variety of ways, though research most prominently focuses on rhythm and/or tapping-based tasks (e.g., Repp & Doggett, 2007; Duke, 1994; Drake, 1998; London, 2012: pp.172). For example, one study from Chen, Penhune, & Zatorre (2008), investigates the neural correlates of musicians' relatively enhanced ability to synchronize precisely timed sequences to musical rhythm (i.e., auditory-motor integration). Chen et al. explored how increasingly complex rhythms might affect synchronous behaviours and neural activities. Participants were balanced for musical training (12 musicians & 12 non-musicians) and for biological sex. In a session subjects heard and were asked to imitate three different rhythms by tapping in synchrony on a computer mouse. Relevant to the present discussion, the musicians were observed to out-perform non-musicians at all levels of rhythmic complexity. The authors explain that, in debriefing, they discovered this differential performance could in part be attributed to different strategies used across groups to organize elements of the rhythmic sequence (i.e., metrical subdivisions for musicians vs. chunking for non-musicians).

Moving on to experienced-based differences in neural responses, the fMRI results suggest that much of the neural activity observed during this experiment was similar across participants despite musical training, however musicians were found to recruit the prefrontal cortex to a greater extent than were the non-musicians. In other words, neural responses observed in this study suggest that differences in performance were not associated with areas of the brain recognized for fine motor control, but instead were found in areas associated with organizing temporal information and working memory function. It appears that musicians have developed the ability to manipulate and compartmentalize timing- and sequence-based information, resulting in differential performance and neural activity that can be predicted by way of previous experience i.e., musical training. These neurological findings support the strategy-based explanation from the authors, and illustrate how experience-based effects might be recognized on both performative and neurological levels.

1.3 Setup, Experimental Scope, and Research Questions

When considering potential convergence to background noise, the above literature indicates that exploring variation in pitch, duration, and intensity would be a sensible place to begin. However, this raises the question of *which type of* noise would be most appropriate for this type of work? Importantly, music and language share these dimensions as the effective building blocks of expression in message transmission. Perceived volume-related dynamics are well-recognized and integral components in both speech and musical composition, and Patel (2008, chapters 2&3) devotes lengthy discussion to comparing the roles of pitch and rhythm in music to their roles in spoken language. In much the same way that musical notes carry different frequency values (pitches) and durations, we see similar variation in spoken language distinguishing potential phonemic and sub-phonemic ambiguities. For instance, many languages of the world maintain some form of pitch-based contrast, and these have been associated with segmental, syllabic, lexical, and phrasal levels. Where languages like Mandarin Chinese utilize a system of tones and tone-contours to distinguish speech sounds and syllables (Wang, 1967; see Yip, 1980 for discussion), most (all?) English dialects exhibit variation in pitch that can alter the meaning of both words and phrases (e.g., Pierrehumbert, 1980; Hirschberg, 1993). Bengali also exhibits distinguishing pitch information at the phrasal level (Hayes & Lahiri, 1991), whereas some Bantu languages are noted for differentiating verb forms using tone/accent (Odden, 1995 pp. 449).

Though changes in pitch and duration can often co-occur in the marking of lexical and sentential stresses (*or pitch accent*), some languages focus primarily on the duration of certain segments to important end. For instance, Māori and Samoan phonemically contrast vowels of relatively long- and short-duration, all other things being equal (Harlow, Keegan, King, Maclagan & Watson, 2009 pp. 131-132; Hovdhaugen, 1992). Durational distinctions are not restricted to vowels however: Finnish and Japanese are examples of languages that show comparable long and short phonemic contrasts with stop consonants in geminates (Kunnari, Nakai & Vihman, 2001; Han, 1992). Indeed, when considering duration as an acoustic variable of importance, phrase-final lengthening of segments has been well documented in many languages and noted as helpful when language learners begin to parse word boundaries and as a cue in turn-taking among other things. Thus, whether these pitch- and duration-related differences in speech are the product of syntactic positioning, pragmatics, or exist as phonemic or prosodic distinctions, it is currently well

documented that spoken language and music share some of the primary building blocks through which acoustic information is transmitted.

Therefore, when investigating whether or not human speakers entrain acoustically to ambient noise, previous work motivates use of background music and exploring potential convergence to pitch, intensity, and rate/tempo as a reasonable starting place. But what are the specific questions that are being addressed by this work? Broadly speaking, the following experiments contribute to better understanding how speech production can be influenced by environmental factors. More specifically, this work investigates two overarching questions:

I. **Are speakers reliably influenced by acoustic characteristics of ambient noise (in this case background music), such that acoustic characteristics of their speech become systematically more or less like corresponding characteristics of that noise?** General tendencies toward convergence have been identified through previous studies on human subjects, and entrainment to both communicative and ambient speech has already been shown. Given the many acoustic dimensions implicated in Pardo's review (2013) on acoustic-phonetic convergence in speech production which are, for the most part, restricted to the context of speech, the acoustic dimensions of Pitch, Intensity, and Tempo/Duration remain as the most sensible acoustic dimensions to test for increasing (or decreasing) similarity over time when considering the context of background music.

And this raises an interesting follow-up question regarding the specifics of any potential acoustic entrainment to ambient noise:

II. **Is there meaningful variation between individuals or groups in terms of convergent and divergent behaviour – and, if so, to what extent might this variation be predictable through social- or experience-based variables?** Previous research has shown that certain forms of convergence take place without any need or influence of agency or social/communicative interaction (e.g., collective convergence, as exemplified through a table full of metronomes: Pantaleone, 2002), though other forms are primarily rooted in social interaction or motivations (e.g., estrous synchrony: McClintock, 1971). Indeed, some forms of linguistic-convergence simultaneously involve both automatic and social components (e.g., Convergence to specific vowels: Babel, 2009; Convergence or divergence to express identity: Drager, 2011), and the relative contributions from these influences in different contexts are not yet known. Thus, if acoustic convergence during speech production is purely a social effect, then convergence to background music would not necessarily be expected to occur. However, if this type of convergence is solely an automatic effect, or simultaneously both automatic and social in drive (a combination of the two would likely be suggested by Babel, 2009), then perhaps different outcomes will be observed through various combinations of social and experiential motivations. In other words, it seems possible that a speaker's past experience, expectations, and attitudes/opinions may also influence their susceptibility to acoustic convergence when presented musical noise in a reading task.

Therefore, aiming to further our general understanding of how environment and context can influence language use, this dissertation investigates the possibility of speakers entraining to specific acoustic properties of ambient noise, primarily encountered in background music. A series of three experiments are described below, all of which contribute to addressing the above research questions.

Experiment 1 serves as both the first step in testing for potential entrainment effects (and subsequently any call for further experimentation on the topic), and also introduces a novel method for stimuli generation through the appropriation of software previously unused in phonetic experimentation. This study employs a speech-in-noise reading task to investigate whether or not speakers entrain to variation encountered in the pitch, intensity, and tempo/duration envelopes as presented in musical background noise. Based on works discussed above, it is predicted that speakers' productions will be reliably influenced by acoustic characteristics of that background noise. Through previous research exploring accommodation in speech production, we know that speakers are known to converge/diverge with certain acoustic properties of their speech partners' productions; a lack of such effects when exposed to background music could therefore be interpreted as support for acoustic-phonetic convergence being an entirely social process, requiring a communicative context in order to take place. Any observed convergent (or even divergent) behaviours related to background music could be interpreted as support for increased levels of automaticity in the mechanisms driving acoustic convergence; however, use of experientially-based information for prediction in analyses provides the opportunity to explore the possibility of both automatic and social contributions jointly, as well as potential contributions from previous experience and expectation. For this and the following experiments, background music is composed specifically to isolate the acoustic dimensions of pitch, intensity, and duration in order to manipulate or focus on a single acoustic dimension at a time.

Experiment 2 serves as a replication of Experiment 1 with a slightly altered methodology, addressing issues recognized in both design and analysis. This study explores only effects rooted in Pitch and Intensity manipulations. And, after recognizing methodological issues which convoluted the analyses of Experiment 1, changes were also made to the reading materials (specifically, a phrase list was used instead of connected speech passages) in order to maximize the number of usable observations and to meet all assumptions of the statistical tests used during analysis. Additionally, alternative statistical methods were explored in this study, resulting in complementary analyses used to inform and support one another. These changes facilitate a more straightforward assessment and a better understanding of the data.

Experiment 3 adopts the design and methods of Experiment two, and narrows the focus to pitch-related entrainment only. However, recognizing potential weaknesses in previous research that played a key role in motivating the present work, pitch-convergence is compared and contrasted in treatments exploring both ambient music and additional ambient speech conditions. Analytical methods are identical to those of Experiment 2, though conditions have been designed to allow for direct comparison of music-based and speech-based effects.

CHAPTER 2: Experiment 1 (Convergence in Speech to Background Music)

2.1 Introduction

Linguistic convergence can be broadly defined as *a situation or context in which some characteristic(s) of language use, between two or more language users, becomes more similar through contact* – though, recall that convergence has been operationally defined within the present work as *a process whereby movement observed for one entity in a particular acoustic dimension influences movement in another entity, such that the two exhibit similar change over time*. Previous research has found strong support for effects of convergence in language use (see section 1.2 for discussion). Acoustic-phonetic convergence has been tested extensively in communicative-speech contexts (e.g., Levitan & Hirschberg, 2011; Pardo et al., 2012; Babel, 2012; Pardo, 2013), though only two known studies explore potential entrainment to ambient speech (Delvaux & Soquet, 2007a; 2007b). In fact, there are no known studies that explore the potential for acoustic-phonetic entrainment to background noise in speech production. The present experiment, therefore, recognizes this important gap in the literature and serves to help better understand some ways in which spoken language may be influenced acoustically by ambient noise.

The present work specifically investigates potential for acoustic convergence in speech production to background noise, tested in three acoustic dimensions using carefully constructed background music. This work serves as the first step in testing whether or not any such effects may be observed outside of the realm of spoken language, and also introduces novel methodologies with regard to stimuli generation that allow for high quality manipulation of polyphonic, polyrhythmic signals in multiple acoustic dimensions (methods are discussed below in sections 2.2 and 2.4).

A speech-in-noise reading paradigm is used in this study, where speakers read passages aloud while background music was played quietly through isolation headphones. Experimental conditions were designed to explore potential entrainment to pitch, musical tempo, and intensity. Stimuli were generated using a single musical piece composed specifically for this work. Many potential confounds were taken into consideration during composition and, as a result, the piece used as a base for stimuli generation maintains relatively flat tonal, temporal, and intensity contours throughout. These acoustic characteristics remain relatively stable over time, allowing for each manipulation to be applied gradually. Gradual manipulations should help control for effects that may persist over the course of a given condition, and should also facilitate a more straightforward recognition of trends during analyses.

2.2 Musical Stimuli Design

Patel (2008: pp. 9-238) describes certain acoustic properties known to affect the dynamics of both spoken language and musical composition, and does so with specific emphasis on pitch and tempo. In order to test for convergence in speech production to acoustic information within a musical signal, one must carefully control for other properties that are likely to vary similarly, and simultaneously within those signals. Choosing stimuli for this type of work is thus inherently difficult because musical compositions tend to vary in multiple dimensions, and rarely restrict this type of variation at a given time point to just one form of change (much like human speech!). Please note that the use of musical stimuli as background noise by nature involves complex signals comprised of multiple simultaneous voices and rhythms, and as a result absolute targets are most often unavailable for measurement. When considering voice-pitch for example, it is not possible to know which musical voice a speaker may converge/diverge with at a given moment because there are typically multiple instruments sounding at that moment (see below discussion in section 2.3.2.2). Moreover, productions may be further influenced in this way by the speaker's previous experience (see discussion in section 2.3.1.1). Therefore, with later analyses in mind, stimuli throughout the present work have been designed to consider relative differences in production as opposed to measuring absolute distances from (largely unknown) targets.

Described below are 7 areas requiring attention when selecting – or composing – music for this type of work. I also provide descriptions of how each concern has been addressed. In presentation below, these issues are broken into two overarching categories: (1) General considerations for musical selection (that is, primarily rooted in psycholinguistic associations given previous experience), and (2) Specific considerations during stimuli manipulation (being related more so to physical/acoustic properties of the signals themselves, and how these properties may affect processing). Recognizing how difficult it would be to find a pre-existing composition that satisfies all criteria, I have elected to compose music which adheres to the following constraints. However, certain difficulties in composition stem from required variation in the characteristics of interest; such variation over time contributes to the perceived musicality of that piece (paralleling the naturalness of speech). Put simply, the underlying aim in composition was to *imply* sufficient change in a variety of acoustic dimensions in order for listeners to recognize the piece as sounding musical, while in fact the stimulus remains relatively consistent in these areas over the course of the piece. The composition used as a basis for stimuli generation will be referred to herein as *Science Music*. The composition was written in collaboration with Rob Batke of *Artisan Loyalist,*[1] and has a run-time of 00:03:27.

[1] The song used as a base for all stimuli was written in collaboration with Rob Batke of *Artisan Loyalist* (http://www.artisanloyalist.com/). *Science Music* was written and recorded using both Ableton Live 9 (Ableton, 2015) and Garageband (Apple Inc., 2012) software.

2.3 Controls and Considerations: *(Decisions made to address each consideration* have been italicized for convenience)

2.3.1 General Considerations for Musical Selection

2.3.1.1 The Influence of Previous Knowledge

Experience and expectation are known to influence many aspects of language use (e.g., Phonetics: Hay et al., 2017; Morphology: Reichle & Perfetti, 2003; Syntax: Jaeger & Snider, 2013; Psycholinguistics: Zwaan, 1994; Phonotactic acquisition: Edwards, Beckman & Munson, 2004) as well as musical perception (Pearce, Ruiz, Kapasi, Wiggins, & Bhattacharya, 2010; Schmuckler, 1989; Patel, 2008: pp. 26-28, 77, 84; Fraisse, 1982; Meyer, 1957; Pearce & Wiggins, 2012). Acknowledging the well-known frequency effects often observed in linguistic research[2], as well as the many similarities between speech and music discussed throughout this and other works, it is possible that familiar or known musical stimuli could distract participants and/or colour their performance in some way, thus contaminating the data. Indeed, as mentioned earlier, previous work has also shown that trained musicians at times exhibit different neurological responses, fine motor control, and organizational strategies than those with no formal musical training (e.g., Chen et al., 2008), further complicating personal experience as a concern. Therefore, when investigating the potential for speakers to converge acoustically with ambient music, it would be wise to select a piece that participants are not well acquainted with for use as a base-stimulus. A fine balance, however, must also be recognized between interest and boredom. Participants cannot be overly engaged by the piece playing in the background, and simultaneously should not be annoyed by redundant music continuing throughout the 5 conditions – two of which are theoretically identical controls, lacking any acoustic manipulation. *Science Music therefore emulates the style of a popular musical group whose general feel and song structures could be adapted to accommodate the necessary constraints, resulting in stimuli that are interesting enough to avoid boring participants (i.e., fatigue effects) while also avoiding high levels of distraction and predictability due to participants 'knowing' the piece*. Moreover, the experimenter should me mindful of predictable differences in performance due to personal experience. Thus, *musical experience has been controlled for in the present work by collecting information about participants' musical training for use in later analyses*.

2.3.1.2 Distraction and Reading Ability

The issue of engagement vs. fatigue is further addressed through attention and the cognitive demands involved with coupling speech production with a reading task. The act of reading requires subjects dedicate certain cognitive resources to that task (e.g., Krainik, Lehéricy, Duffau, Capelle, Chainay, Cornu, Cohen, Boch, Mangin, Le Bihan & Marsault, 2003; Cucchiarini, Strik, & Boves, 2002), though not all readers do so with the same level of

[2] '*Frequency*' in this context refers to how often some thing or concept is or has been encountered, as opposed to physical vibrations or periodic waves in the environment.

skill. *The passages selected for this work however, while not overly demanding, exhibit a style and vocabulary appropriate for university level talkers*. These reading materials were chosen for a specific level of complexity, to draw on cognitive resources such that the introduction of background music is noticed by speakers, though remains peripheral as a draw for attention in comparison to the actual speech production.[3] And because this joint-task involves resources beyond those required for speech production alone, it is believed that specific properties of the background noise are less likely to be focal for participants in this context.

2.3.1.3 Linguistic Information Encoded in "Vocals"

It is further possible that the inclusion of overtly linguistic information (i.e., vocal melodies, lyrics, cultural associations, etc.) could influence speech production, and that this could happen in both active and passive ways. Drawing on a large body of sociolinguistic literature, previous studies describe speakers' production patterns changing to reflect group membership (e.g., Labov, 1972: pp. 43-54; Hay & Drager, 2010; Drager, 2011) and perceptual boundaries shifting due to associations between sung dialects and background music (Gibson, 2008). Indeed, other work shows fine acoustic detail in the speech of bilinguals changing in both languages to reflect properties of the ambient language (Sancier & Fowler, 1997). Though the speaker is not always aware of (or choosing to make) such changes, overtly linguistic information must be avoided as a potential confound, at least until a baseline has been set regarding phonetic-convergence to ambient noise. *Science Music, therefore, was composed as an instrumental piece in order to minimize the possibility of influencing participants in unintentional ways through the inclusion of overtly linguistic information*.

While Science Music does include synthesizer-generated lead lines over the chord progression where a vocal melody would normally be situated in the mix, these "voices" are clearly synthetic (that is, generated with synthesizers as opposed to the more organic sounds which might be generated using acoustic instruments or the human vocal tract); as such, these voices are relatively unlike natural speech. The melodies included in Science Music have been designed to fill two explicit roles: (1) Recognizing the relatively simple chord progressions and song structure dictated by harmonic constraints (more on this below), *melodies have been designed to add sonic textures over repetitive structural segments (e.g., VERSE, CHORUS, etc.) in order to break up monotony for the listener, taking the place of a vocal melody without being overly engaging or speech-like*, and (2) Again recognizing the underlying simplicity of the composition, *these lead lines have been overlaid asymmetrically – with some leads continuing into the section that follows – such that the VERSE+PRECHORUS+CHORUS structure of the chord progression is less apparent*. While risking a more interesting composition overall, the addition of melody in this way seemed a reasonable concession in order to avoid participants being distracted by monotony, annoyance, or fatigue.

[3] Complexity of the selected reading materials was determined in a brief pilot study, where participants read several passages of variable difficulty under test-like conditions and discussed their level of distraction given the music in conjunction with the reading material.

2.3.2 Specific Considerations During Stimuli Manipulation

2.3.2.1 Intensity

Much like spoken language, music rarely maintains constant intensity for long periods of time. Intensity-based variation creates a dynamic range that draws attention to particular rhythmic and melodic elements of importance, and can also be used to convey emotional intensity through perceived changes in loudness. Though realized in different contexts, the function of intensity-based dynamics in music is often quite similar to a speaker contrasting phonemes (or even ideas) in speech production. With this similarity in mind, it would be wise to consider how quickly listeners and speakers may recognize and be affected by variation in the intensity envelope, as well as the length of time that any such effects may persist.

For example, Howell (2008) suggests that certain well-known compensatory mechanisms such as the Lombard effect (Lombard, 1911) – where speakers alter vocal intensity based upon relative environmental noise levels – and the Fletcher Effect (Fletcher, 1918) – where speakers alter vocal intensity as a function of attenuated or altered vocal feedback levels – exist as the result of a negative feedback mechanism. One must also consider how quickly (or slowly) these types of compensation may be realized (i.e., Are compensations instantaneous? Or is there a considerable lag following the presentation of a stimulus before compensations in production are maximized?). Heinks-Maldonado & Houde (2005) give some insight on this last issue, describing significant latencies in compensation following unexpected variation in vocal loudness feedback, where delays of similar mean duration were generally observed across participants. The study describes mean latencies (as measured from perturbation onset to response onset) in the order of 171 ms when feedback-intensity was lowered by 10dB, and roughly 287 ms when signals were increased by 10 dB, suggesting that the specifics of compensations were at least partially contingent on whether a signal's intensity had been increased or decreased. It appears, then, that speakers are relatively slower to lower their speech levels than they are to raise them in response to manipulated feedback levels.

Acknowledging the well-documented prevalence of convergence in humans, and the potential for intensity-related effects in speech production to persist over time (e.g., Goldinger & Azuma (2004) describe effects persisting for as long as a week), it becomes necessary to carefully control for variation in the intensity envelope when exploring the possibility of intensity-related convergence. Unfortunately, music lacking intensity-related dynamics is most often recognized as uninteresting and lacking musicality. The present work therefore, required a base-stimulus with relatively little variation in the way of intensity that could simultaneously maintain perceived musicality. Further complicating this problem is the fact that, when exploring potential entrainment to intensity, manipulations must be applied so that any observed changes in speech production are attributable only to variation in the intensity envelope and are not simply artifacts of other well-known effects associated with noise masks of sufficient volume, such as the Lombard effect (cf. Lombard, 1911; and see Brumm & Zollinger, 2011 for a review).

To mediate these intensity-related issues, *dynamics in the way of perceived volume/intensity were implied by adding and removing various instrument voices and melodies over the course of the piece*. For example, from 00:15-00:59 (verse 1) instrumentation remains relatively consistent; however, as Science Music nears the one minute mark an additional synthesizer is introduced to fill more sonic space. This change both increases instrumentation

and, as a consequence, produces the effect of an increase in perceived volume. It should be noted that under typical circumstances this type of addition may actually result in an increased amplitude of the overall signal. Therefore, *a strict compression was applied to the Master bus (mixer channel) prior to exporting the composition in order to minimize variation in the intensity envelope*. When describing how this process (i.e., compression) affects a signal, the Ableton Live 9 manual (DeSantis, Gallagher, Haywood, Knudsen, Behles, Rang, Henke, & Slama, 2015) explains that:

> "A compressor reduces gain for signals above a user-settable threshold. Compression reduces the levels of peaks, opening up more headroom and allowing the overall signal level to be turned up. This gives the signal a higher average level, resulting in a sound that is subjectively louder and 'punchier' than an uncompressed signal." (pp 301)

Thus, the louder points in a signal can often be brought down to match the level of quieter portions of the same waveform and, as a result, the overall level of the newly compressed recording can then be raised by globally increasing gain to a relatively higher level without fear of digitally clipping those peaks. Using compression in this way is especially useful in the present work, as it provides the opportunity to generate waveforms of relatively constant amplitude while simultaneously allowing for implied intensity-related dynamics through varied instrumentation.

While it is possible that speakers may converge to 'implied intensity' in a way not unlike convergence to actual intensity-related changes, I do not know of any work that tests this possibility explicitly. That is, known previous works consistently focus only on actually increasing or decreasing signal presentation levels. Having considered this possibility though, I believe that the likelihood of convergence to implied intensity should be considerably lower than the potential for convergence to actual signal level manipulations. Conversely, if Science Music stands out to participants for lacking musicality due to insufficient intensity-related dynamics, this would likely draw increased attention to the background 'noise' and could raise further complications through directed attention. Finally, the intensity-based signal manipulation in the present work (described in detail below in section 2.4) is rather dramatic, so effects related to actual signal manipulation are likely to far outweigh any possible convergence to the more subtle changes which might occur through implied-intensity. A hybrid method of implied dynamics and signal compression seemed the best available option, and the possibility of convergence to implied intensity will not be discussed further.

2.3.2.2 Pitch, Tonal Centre, and F0

A second point requiring consideration is the clarity of a tonal centre, or what can be conceptualized as a 'home frequency' that speakers might gravitate toward (cf. Auditory Scene Analysis, see Bregman, 1994 for in depth discussion). The notion of home frequency will be used in the present work instead of a specific frequency at times, and can most readily be compared to the musical tonic (or 'root'*:* the first note in the scale) for the key in which a

song has been composed. Presuming listeners do in fact entrain to pitch-variation, I propose that they may do so to a home frequency, where speech follows changes encountered in musical voices/instruments/melodies that have been subjectively attended to as part of the complex musical whole. In other words, different people may entrain to different components within a musical piece (cf. the description of convergence as generally inconsistent in Pardo, 2013).

However, the proposition is complicated slightly by the fact that music is most often constructed using multiple notes, voices, and melodies simultaneously (i.e., polyphony), where many frequencies and harmonics are concurrently present. The proposed concept of home frequency is therefore not restricted to a particular octave, or even a specific degree of the musical scale, but instead considers information regarding the diatonic notes within a key more generally. In this way, speakers may converge with any voice that they subjectively recognize to be most salient.[4]

With this thought in mind, one way to test for convergence in voice-pitch would be to *use a song that maintains a clear key/home frequency throughout as a base for stimuli generation. Such consistency can be achieved relatively easily through a composition that does not modulate (i.e., change keys) or borrow chords from other keys. Further contributing to such clarity/salience are the chord progressions throughout Science Music, which can be described as relatively consonant* (i.e., deemed 'pleasant'; Helmholtz (1954: pp. 182) describes the effect of consonance as something listeners relate to a greater coincidence of upper partials). This quality is often associated with simpler chord voicings, drawing directly from the (diatonic) notes available in a chosen key and typically maintaining sufficient distance between the notes involved in a chord's construction. Contrasting consonance, dissonant voicings typically imply required resolution and are often used to create musical-tension – a tool often used to change musical keys. In fact, dissonant voicings have even been noted for evoking fear in the listener at times (Trainor, 2008). Adhering to the above constraints will clearly indicate a home frequency to the listener, providing a sufficient base from which to explore entrainment to a song's tonal centre through global pitch manipulation.

Before moving on, one should note that Bauer, Mittal, Larson & Hain (2006) describe listeners' response latencies to changes in pitch as similar to those described above for intensity. With this knowledge in mind, *all manipulations to pitch/home frequency have also been applied in a similar linear fashion over time*.

2.3.2.3 Rhythm, Consistent Tempo, and Speech Rate

There is some evidence suggesting that speakers may converge in speaking rate with other talkers (e.g., Putman & Street, 1984; Manson, Bryant, Gervais & Kline, 2013; Schweitzer & Lewandowski, 2013). Therefore, another potential problem lies in variation often observed in musical rhythmicity – though, here I speak only to the intentional, and often abrupt, changes in tempo, time signature, and general 'feel' found in certain musical styles and

[4] Here, I am allowing for subdivision of both the available instrumentation and the chords potentially played by each of those instruments when referring to the 'voice' a listener finds most salient.

exclude the more gradual, and often less intentional speeding-up and slowing-down encountered in some live recordings and performances. Certain artists and ensembles choose to switch genres, tempos, and time signatures within compositions or even within sub-sections of songs (e.g., Rush: "Limelight (1981); Dirty Projectors: "Offspring are Blank" (2012); John Zorn: "You will be Shot" (1990); Candiria: "Without Water" (2002)). Such change can be a useful compositional tool when aiming for specific effect, though it can also alienate and be confusing for some listeners. If nothing else, this type of variation robs a composition of a consistent pulse, forcing the listener to work harder cognitively and recalibrate meter and tempo with every change. Therefore, in order to most clearly identify any potential effects that musical tempo may impose upon speech production, a single unambiguous pulse and time signature should be maintained throughout all conditions to allow for gradient manipulations in timing, which will facilitate more reasonable inferences in this regard.

Thus, Science Music maintains a single musical style (dance/indie-pop) for the song's duration. Moreover, in the way of rhythm, variation in feel & tempo (or groove) within Science Music is once again implied through alternating kick drum patterns while both the snare/claps and high hat pattern effectively persist throughout. Structuring rhythmic elements in this way consistently places appropriate percussive voices on the off-beats of beats 1 through 4 as well as on the beats that typically drive rock and pop music (i.e., beats 2 and 4), known as a 'back beat'. Though the kick drum patterns do allow for some relatively interesting variation in the overall scheme of the composition, consistency of the snare/claps and high hat pattern very clearly indicate an unwavering global tempo that is clear to the listener from the song's onset to completion. Much like the instrument changes related to perceived intensity-dynamics described above, variation in the kick drum patterning seemed a reasonable allowance in order to avoid monotony and retain perceived musicality for the listener.

2.3.2.4 Preferred Tempi (and Generalizability)

Additionally, there are tempo-related issues one must consider when approaching this type of work. For example, a variety of studies suggest human subjects exhibit regular, and preferred rates for general motor actions (e.g., Harrison, 1941; Smoll, 1975; Smoll & Schutz, 1978; Taguchi, Gliner, Horvath, & Nakamura, 1981; Farnsworth, 1950: pp. 54-56). Such preferred tempi have also been noted for activities more straightforwardly related to music, like dancing or tapping to a beat (e.g., Moelants, 2002; 2003; Patel et al., 2009; Farnsworth, Block, & Waterman, 1934; Repp, 2005; and see Fraisse, 1982 for in depth discussion). LeBlanc & McCrary (1983) explain that when presented jazz music of progressively quickening tempi, upper elementary students showed significant preferences for the faster music with each increase. Fraisse (1982) describes reasonable inter-subject variation, though suggests a preferred tempo in the neighborhood of 100 beats per minute (bpm) would generally be most representative. However, having subsequently analyzed an extensive corpus of popular music, including different styles from different time periods, Moelants (2002) argues the majority of popular/preferred music exists between 120 – 130 bpm, and further notes that the optimum range for musical tempi observed in his data (being between 81 – 162 bpm) very closely resembles the optimum range for average walking speeds (81 – 150 steps per minute). Taking into account the range of tempi implicated in previous works *Science Music was composed at the maximum arguable*

preferred tempo of 130 bpm, anticipating a decrease during the duration/tempo condition to a value which remains within the preferred range – thus, marginalizing the likelihood of any potential effects in speech rate being driven by gravitation toward other preferred tempi, or potential false negatives due to manipulation outside the preferred range.

2.3.2.5 Potential Dichotic Listening Effects

Finally, it has been suggested by Berlin, Lowe-Bell, Cullen, Thompson, and Loovis (1973) that any acoustic event that can be perceptually linked to rapid gliding motions of the vocal tract may result in a right ear advantage in processing speed. And, though Broadbent (1954) also describes a right ear advantage when processing speech sounds, Kimura (1964) found a left ear advantage for non-speech sounds (e.g., melodies), which calls into question exactly which acoustic information listeners will find most salient in the present context given music's unique position as simultaneously speech-like and noise-like. Recognizing the nature of the stimuli in this work, diotic (one waveform delivered simultaneously to both ears, commonly referred to as 'mono' delivery) vs. dichotic delivery (where a different waveform is simultaneously sent to either ear, commonly referred to as delivery in 'stereo') could reasonably result in differential effects due to available cognitive and physiological resources, processing, and attention. *Original waveforms were therefore generated as both Mono and Stereo versions before applying any acoustic manipulation* in hopes of eventually investigating this potential for response differences, maximizing comparability of stimuli across presentation-conditions. However, in order to keep the task manageable for participants it was decided that only diotic stimuli presentation would be used in all treatments for the present work. Stereo manipulations will therefore not be discussed further, though they have been generated and are available for use in later experimentation.

2.4 Stimuli Generation:

A brief note on terminology: Within the present works I will use the term *Production-stimulus* when referring to the items presented visually to speakers as prompts to be read aloud; the term *Stimulus (or, Background-stimulus)* is reserved for different forms of background music/speech presented auditorily during treatments (i.e., variants of Science Music, or in EXP.3 multi-talker babble).

2.4.1 Background Stimuli

Having composed Science Music with the above criteria in mind, three manipulations were applied to the Original 'mono' recording. Such manipulation resulted in four conditions: (1) Original/unaltered, (2) Intensity, (3) Duration, and (4) Pitch.

Manipulations for each treatment involve one of the three acoustic variables (i.e., intensity, duration, & pitch) gradually deviating from a relatively static baseline over time and then, after reaching a specified target,

returning to the original value in the same linear fashion. Manipulation in two directions seemed a sensible approach, as speakers gradually increasing or decreasing the intensity, pitch, or rate of their productions over time could occur as the byproduct of any number of factors (e.g., insufficient subglottal pressure, fatigue, excitement, etc.). If deviation occurs naturally for any reason, it should be recognizable in the analyses of control conditions, where participants experience music lacking any acoustic manipulation. However, if productions increase and then decrease (or decrease and then increase) over time as per the modified treatment-envelopes, then such effects would certainly support speakers as being influenced by the stimuli. By this reasoning, manipulations in EXP.1 deviate and then return to the point of origin in all test conditions.

In order to differentiate effects driven solely by the introduction of music from any effects more directly related to the acoustic manipulations described above, the initial 30 seconds of each background stimulus was left unaltered in all conditions and can be treated as a baseline for analyses as well as an acclimatization period allowing speakers to deal with any potential surprisal-effects that come with the introduction of music. Manipulations are, therefore, applied from the 30-second mark until the song's completion. The Original condition serves as another form of control and remains unaltered for its entirety; this lack of manipulation allows for recognition of any longer-term effects of speaking in musical noise, and also sets a secondary baseline from which to measure potential variation within conditions over time, should the analysis become less straightforward.

Background music for the Intensity condition was altered using Praat (Boersma & Weenink, 2014) by generating an Intensity tier for the .wav file of science Music and then retracing the envelope over time. Breakpoints were placed at the 30-second mark, the end of the song (207 seconds), and at the mid-point of the manipulation area (118.5 seconds into the composition). Where the first and final breakpoints remain consistent with the file's original intensity level, amplitude of the waveform was increased by +6 dB at the 118.5-second mark. Therefore, all samples on either side of the center target-value were subject to gradual increase/decrease between the file's original level and that of that target, imposed as a linear function (as seen in Figure 2.1). The manipulated envelope was finally merged with the Original source file by selecting both items in the Praat Objects window and using the *Multiply…* command.

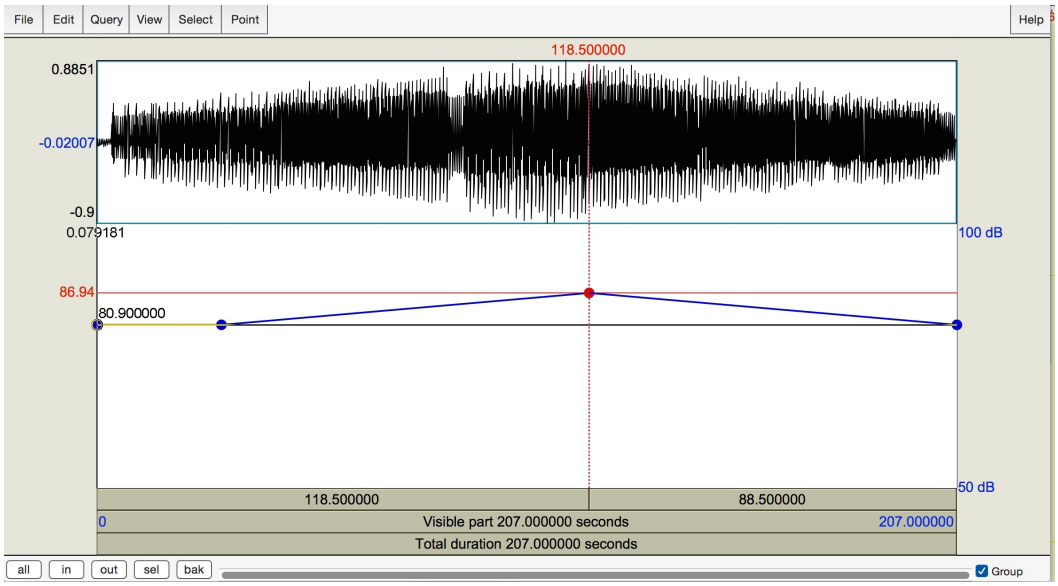

Figure 2.1 Setting breakpoints and manipulating the Intensity envelope in Praat.

Praat struggled to deliver equally high quality manipulations for the Pitch and Tempo conditions. As a result, these treatments were generated using Ableton Live 9 (Ableton, 2015), software designed specifically for creating and manipulating music. Both manipulations were applied via processes similar to that described in the Intensity condition i.e., by placing breakpoints at the same time points outlined above, then retracing the envelope for progressive and linear application in two directions. However, in the Pitch condition the tonal centre/home frequency was lowered by 200 cents (see Figure 2.2) using the Transposition Modulation function with the Complex Pro algorithm; and in the Duration condition global tempo was lowered by 20 beats per minute using the Warp function. Notably, the duration manipulation generates a new sound file with the same number of metrical bars – however, given the variable bar-length that comes with altering tempo, the overall duration of this file was increased to a run-time of 00:03:42.

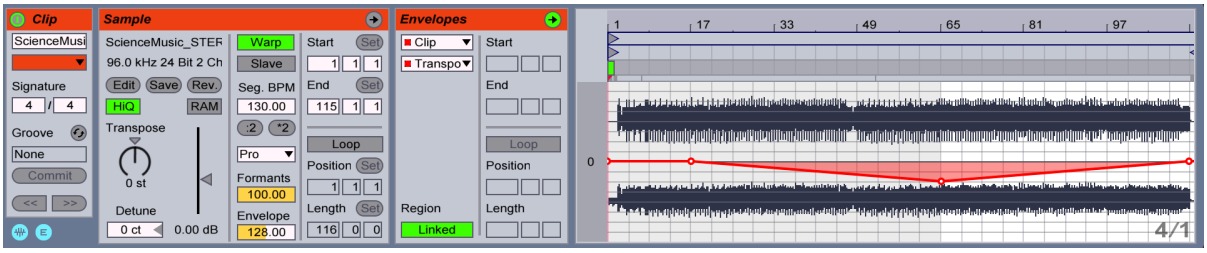

Figure 2.2 Setting breakpoints and manipulating the Pitch envelope in Ableton Live.

Because Live is proprietary software, little information is available about the specific functions and algorithms it might use in different contexts.[5] All stimuli were therefore subject to post-manipulation acoustic analyses (using Praat) to confirm Live had applied manipulations only as directed and not altered signals in other unknown or unintended ways that could potentially influence participants. With these concerns in mind, textgrids were created and coded by hand for all signals, marking intervals (within an interval tier) at each bar line within the song – that is, every four musical beats.[6] Scripts were then composed (also in Praat) to automatically extract bar Duration, Mean Pitch, and Mean Intensity values for each interval (musical bar). Scatterplots illustrating acoustic information by bar/interval for each stimulus are available as Figures 2.3-2.5 (below) and suggest that, admitting some minor imperfections, Live has done a reasonable job of maintaining other acoustic information while applying the intended manipulations. In all figures black triangles represent altered signals and red circles represent the original *baseline* signal.

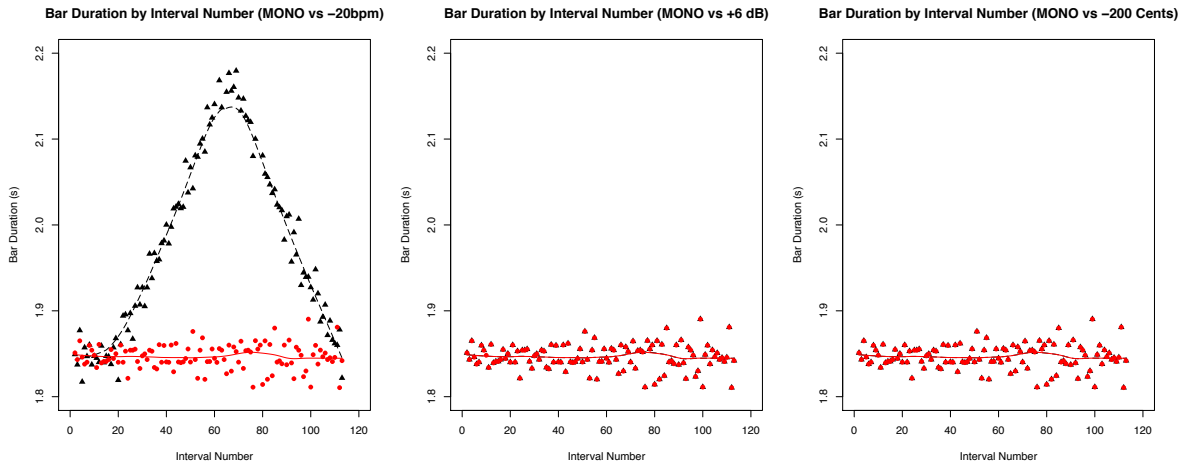

Figure 2.3 Bar duration (on the Y-axis) plotted by interval number (on the X-axis) for each condition. Variation is observed only in the Duration condition, and occurs systematically and linearly as intended. All other conditions exhibit relatively constant interval duration, given hand-coding.

[5] I did contact Ableton with specific questions about how algorithms work but, as one might expect, the company was unwilling to give out information about how their software had been developed.

[6] In fact, only two textgrids were created and these were renamed/reused as appropriate. While the Duration condition required a separate textgrid to accommodate variable bar length, all other stimuli should theoretically maintain a consistent bar duration and, thus, could share the same original textgrid. These assumptions were checked manually and confirmed by a trained phonetician and musician.

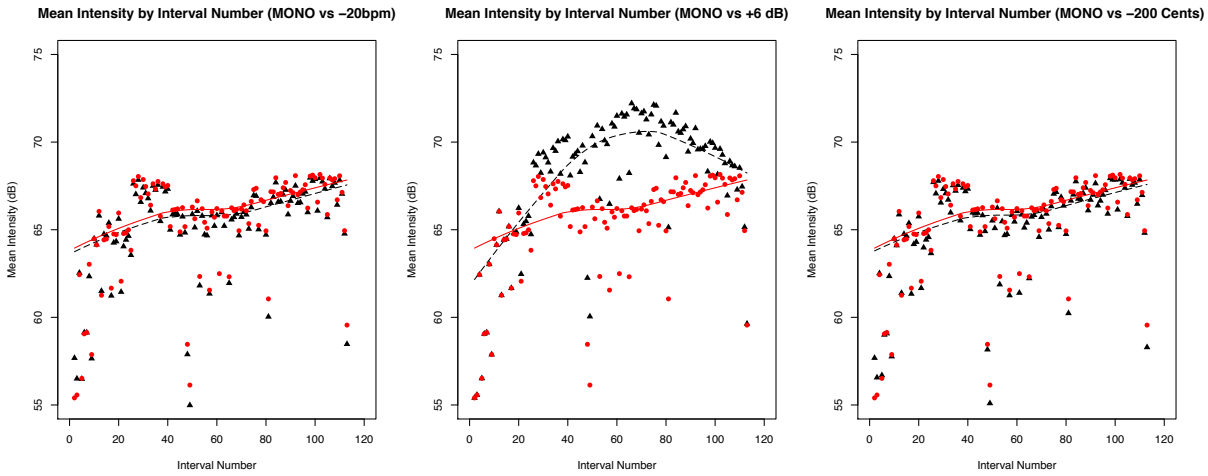

Figure 2.4 Mean intensity (on the Y axis) is plotted by interval number (on the X axis) for each manipulated signal. We find a slight deviation in all Ableton-generated conditions from the baseline signal of roughly -0.5 dB, though this difference is marginal and not expected to adversely affect listeners. The Amplitude manipulation condition, however, exhibits the predicted rise and fall in intensity, where this general trend is not observed in the non-intensity-manipulated conditions.

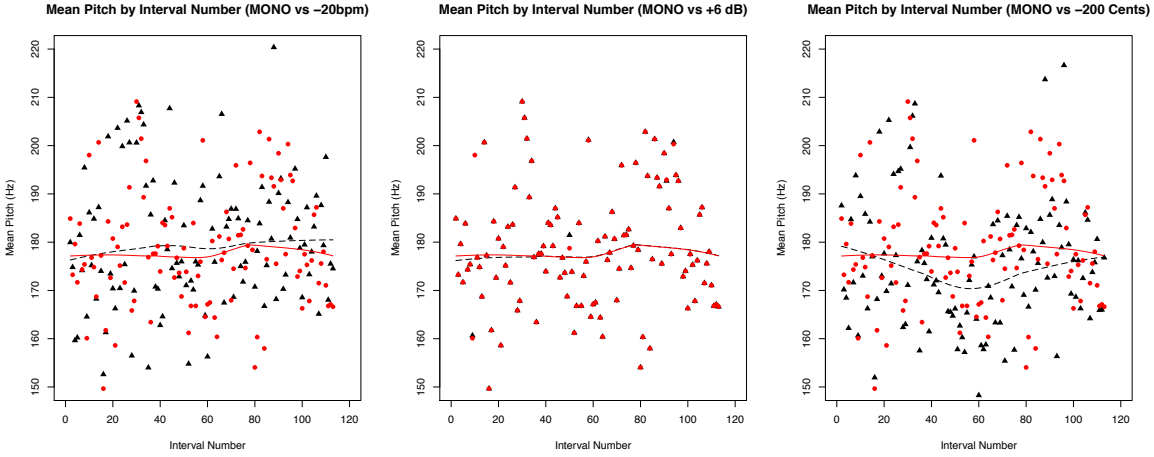

Figure 2.5 Mean pitch (on the Y axis) is plotted by interval number (on the X axis) for each manipulated signal. We see minor variation in all conditions, which was predicted based on how the algorithms can manipulate samples; this is not expected to affect participants adversely. In the Pitch-manipulation condition we see the systematic falling then rising patterns that would be expected given the manipulation, and this pattern is absent in all other conditions.

This confirmation of Ableton's capabilities is reassuring as it supports the introduction of a new and powerful tool for acoustic manipulation in phonetic enquiry. The precision and speed with which Live analyzes and alters complex waveforms, while allowing for both polyphony, polyrhythm, and stereo recordings (which Praat often doesn't deal well with), could make this program an invaluable tool when designing auditory stimuli. Moreover, the

relatively low demands imposed upon the user while generating such high quality manipulations further make this software extremely usable in this context, should one be willing to permit some minor (largely imperceptible) artifacts.

Having discussed the Lombard effect above, it should be noted that Lazarus (1986) describes the known threshold for inducing Lombard speech as ~45 dB(A) for ambient noise and ~55 dB(A) for background speech. Recognizing the many similarities between music and speech, it is currently unknown whether participants would process musical stimuli in this context as more speech- or noise-like. Therefore, background stimuli were delivered binaurally at ~45 dB(A) to all listeners/speakers in order to (1) minimize the likelihood of changes in production driven by the well-known Lombard effect, and (2) explore thresholds which may provide some insight as to how musical noise is processed cognitively. Though the dynamic range of Science Music was compressed to minimize variation in the intensity envelope, the contour is not absolutely flat. Therefore, a relatively consistent (averaged) presentation level was calibrated using a Brüel & Kjær model 4100 Sound Quality Head and Torso Simulator (Figure 2.6).[7] Calibration was based upon mean intensity levels in the Original signal and a slow time weighting (L_S = ~ 45 dB(A)).

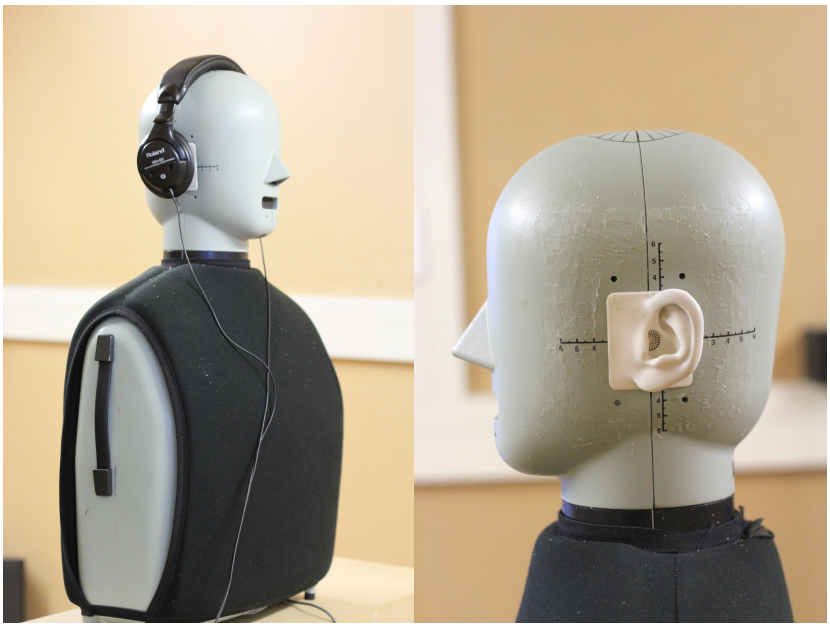

Figure 2.6 The Brüel & Kjær model 4100 is shown on the left 'wearing' the Roland RH-50 headphones used by all participants. The right-most portion of the figure illustrates how the this equipment receives acoustic signals through a synthetic ear canal.

Finally, recall that the first 30 seconds of each signal is not subject to any intentional manipulation. To ensure equivalent presentation levels at the onset of each condition all signals were scaled and matched to 5 decimal points of a Pascal using RMS amplitude values extracted from the initial 30-second portion. All scaling was executed in

[7] I would like to extend enormous gratitude to Ben Scott for his time and help in this calibration.

Praat using the *Get root-mean-squared…* function to extract RMS amplitude values for the first 30 seconds of each file. From these averaged values the smallest was selected as a base for scaling (to avoid clipping), and all other values were then manually divided by that base-value in order to find the factors by which they should be increased using the *Multiply…* function. This process was much like that described above when generating the intensity treatments.

All background stimuli used in this experiment are available online for download through the following link: https://github.com/RyanPodlubny/ScienceMusic

2.4.2 Production Stimuli

Participants read passages chosen from National Geographic magazine while music played in the background. The complexity of selected reading materials was determined in a brief pilot study, where participants read several passages of variable difficulty under test-like conditions and discussed their level of distraction given the music heard in conjunction with the reading material/task. Subsequently, five passages were selected for similar content, style, and difficulty. All passages were altered slightly for ease of reading, to standardize pronunciations, and to remove question marks, exclamation points, and quotation marks – all of which are well known to affect the acoustic variables of interest in speech production.[8] Reading materials were displayed on a laptop computer monitor as PDFs with two-centimeter margins, a justified 16-point Times font, and 1.5x spacing.

2.5 Procedure:

At the beginning of each session participants were seated within a sound attenuated booth and assigned a participant code which includes pertinent metadata – specifically, each code included the experiment name, the year in which the session took place, the participant's chronological rank in the experiment, the participant's (identified) gender, and degree of musical training (e.g., *MMC15_01_F_M* would indicate the first participant in this experiment for 2015, who identifies as female, and is a trained musician). Following informed consent all data and opinions

[8] In all passages quotation marks were removed; question and exclamation marks were replaced with full stops; symbols were replaced with words (e.g., '°F' replaced with 'degrees fahrenheit'); paragraph breaks were removed to avoid utterance-final effects; parentheses were removed and replaced with nothing, or commas if needed for clarity; family names of professionals were exchanged for given names if family names may be difficult for English speakers to pronounce (or variable in how they might be produced); and all numbers were spelled out as words for consistency.

Original passages can be found at:
1. http://ngm.nationalgeographic.com/2013/04/manatees/white-text
2. http://ngm.nationalgeographic.com/2013/12/cougars/chadwick-text
3. http://ngm.nationalgeographic.com/2014/01/komodo-dragon/holland-text
4. http://ngm.nationalgeographic.com/2007/01/humpback-whales/chadwick-text/1
5. http://ngm.nationalgeographic.com/2007/07/birds-of-paradise/holland-text/1

collected were associated only with this number. Once the code had been generated and assigned participants completed an extensive language background survey (available as Appendix 1). Participants then sat a full hearing screening using an Interacoustics Clinical Audiometer, model AC33, which is used to test for receptive sensitivity to a variety of frequencies at various intensity levels in both ears. This screening tests for sensitivity to pure tones at 250, 500, 1000, 1500, 2000, 4000, 6000, and 8000 Hz, where the threshold for 'normal' hearing was defined as a maximum of 20dB(HL). Hearing screenings were included as a more reliable alternative to self-reporting known hearing pathologies that could influence performance, however participants were also asked to report any known hearing impairment. After completing the screening the test and results were explained to each participant and a hard copy of each test was kept for reference and paired with the completed language background survey. Should results have brought to light cause for concern, that participant was referred to the Audiology department at the University of Canterbury for further assessment (n = 1) and excluded from all analyses. This concluded the pre-testing portion of the session.

Participants were then fit with a Beyerdynamic Opus 55.18 MK II head mounted condenser microphone and Roland RH-50 headphones. Signals were routed through a Sound Devices USBPre 2 audio interface and recorded on a late-2013 Macbook Pro laptop computer via Praat (Boersma & Weenink, 2014) at a sampling-rate of 44.1 kHz and bit-depth of 16. This setup allowed for binaural stimulus delivery to the participant, real-time voice-feedback for the speaker with zero latency and, importantly, the recording of stimuli to one channel in a stereo file and the speaker's productions to the other, which are automatically time aligned as a result.

The experiment was run as a speech-in-noise reading task. Reading materials were presented to participants on a Hewlett-Packard EliteBook 8570p laptop computer monitor. Reading was self-paced, and participants were instructed to navigate passages using arrow keys on the laptop computer. Speakers read for ~30 seconds before the introduction of music, which was manually launched by the experimenter. This section where participants read aloud without background noise – in combination with the unaltered initial 30-second portion of each stimulus – facilitates the recognition of any more general effects that may be attributable to the introduction of music from relative silence (as opposed to stimulus manipulation).

The sequencing of production stimuli throughout the study was counterbalanced by subject, aiming to control for any unknown influence of the reading materials and/or the order in which passages were presented. Similarly, condition order was also randomized by subject aiming to control for effects that may be observed as a result of the order in which treatments were encountered. Recognizing the potential for manipulation-related effects to persist over time, and further acknowledging that we are currently unable to predict exactly how long any such effects may persist following a condition, all *test-conditions* were separated by control-conditions, and *all treatments* were separated by a break of roughly two minutes. The experiment was therefore completed in five blocks with breaks of roughly two minutes in between each condition; these breaks were intended to minimize the possibility of any treatment-related effects persisting over time (see Table 2.1 for the general experimental schema). Finally, participants were not told how many conditions comprised the experiment, nor were they told at any time how many conditions remained. This choice was made to avoid rushed or otherwise altered speech toward the end of a session.

Participants were informed at the beginning of each condition that music would begin at some point in the trial, and that they should continue to read aloud until the music stopped.

Condition	Session v.1	Session v.2	Session v.3	Session v.4	Session v.5	Session v.6
Test	Pitch	Pitch	Dur.	Dur.	Amp.	Amp.
Control	Orig.	Orig.	Orig.	Orig.	Orig.	Orig.
Test	Dur.	Amp.	Pitch	Amp.	Dur.	Pitch
Control	Orig.	Orig.	Orig.	Orig.	Orig.	Orig.
Test	Amp.	Dur.	Amp.	Pitch	Pitch	Dur.

Table 2.1 This table illustrates the alternation of experimental and control conditions, where each column indicates one possible randomized order for the presentation of conditions.

Following the experiment participants completed a short, spoken debriefing survey. This secondary survey confirms select information from the language background survey, further addresses known hearing pathologies, level of engagement with the *style* of Science Music (i.e., is this a musical genre you would normally choose to listen to), average time spent listening to music in a given day, whether or not the participant typically listens to music while attending to cognitively demanding tasks (e.g., reading or writing academic papers) and, finally, whether or not the participant consciously noticed any change within the stimuli (of 31 total participants, 5 reported noticing change in the signals).[9] These responses were collected for use as predictors in later analyses.

2.6 Participants:

A total of 31 native and L2 speakers of New Zealand English participated in the study, all of whom were recruited via posters displayed on campus, adverts through various forms of social media, and brief discussions in entry-level linguistics and music courses. Of these participants, 23 provided usable data. That is, 8 in total were excluded for speech and hearing pathologies, reading disorders, or insufficient reading skills in English. Assessments of reading skill were informal; decisions regarding exclusion on these grounds were made by the experimenter on an impressionistic basis after noting remarkably slow speech rate and perceived struggling with the materials. In exchange for their time, participants received a $10 NZD voucher for use at a local shopping centre. Crucially, previous work on multilingualism would suggest that L1 and L2 speakers may perform differently, and that any effects for L2 speakers may be specifically influenced by a relatively increased cognitive load during the combined reading/production task (e.g., Keysar, Hayakawa & An, 2012; Shaw & McMillian, 2008; Geva & Ryan, 1993)[10]. For

[9] 5 participants reported that they recognized changing stimuli. All were asked to expand upon what they thought they heard, and 2/5 gave descriptions of change related to manipulations within the experiment.

[10] Also, see a review chapter from Perfetti (1999: pp. 167-208) describing the cognitive mechanisms involved in reading, which are likely relevant given L1/L2 processing differences. Specifically, see sections 6.7.1 on *Problems in*

this reason, only native speakers will be included in the analyses that follow (n = 15 native speakers). Data from the L2 speakers have been reserved for future exploration, and will not be referred to further.

Multiple subgroups were considered within the sample to explore other related areas of interest with a range of predicted effects. Of the participants whose data will be used, 9 were female and 6 male. Female speakers have been recognized as more susceptible to the (arguably) related Lombard effect (Egan, 1972) and the literature reviewed above describes other gender-based effects in entrainment-specific works; therefore, we may also expect gender-based differences in performance during the current experiment. The sample also considers formal musical training as a predictor. Professional musicians have been trained to recognize subtle acoustic variation (both consciously and unconsciously), and further motivation comes from research showing that musically skilled subjects often perform differently than unskilled participants, most often shown in rhythm and motor-control based tasks (e.g., Duke, 1994; Drake, 1998; London, 2012: pp.172). In the present work, musicians were recognized as belonging to one of three groups: (1) Non-musician (< 6 months formal training in life), (2) Musician (> 6 years formal training in life, AND/OR currently averaging no less than 10 hours per week performance/practice time); (3) Some musical experience (this is effectively an 'else' category which includes participants who fall in between the two more extreme categories). Because trained musicians are skilled at recognizing low-level acoustic variation, and have the ability to quickly – and at times automatically – adjust to low-level variation in tempo, pitch, and intensity while interacting musically with other players, two competing hypotheses make it difficult to predict exactly how training and experience might influence production for these speakers:

1. Given that a great deal of low-level variation in musical performance is accommodated automatically via entrainment (e.g., a bass player synchronizing with an inconsistent drummer, or a vocalist matching pitch to an out-of-tune guitar player), it is possible that musicians would be relatively more susceptible to any potential effects, and potentially *unaware* that they are altering their performance.
2. It seems equally possible that, given such extensive training, musicians may be more likely to *consciously* recognize these low level inconsistencies. As a result, musicians may then resist entrainment to the manipulations due to maintenance strategies (e.g., Repp & Doggett, 2007). If conscious recognition is in fact the case, it seems plausible that effects consequentially might be negated or even inverted. And with specific regard to the pitch condition, it seems possible inversion could result in increased speech intelligibility; divergence in this way would involve frequency-based information becoming maximally different from the noise-mask, a skill possibly developed through experience trying to make oneself understood when regularly competing with music in the speech environment.

---

lexical orthographic-phonological processes, 6.7.2 on *Problems in processing lexical meaning*, and 6.7.3 on *Problems in processing syntax*.

In summary, data from the 15 usable participants allow for exploration of effects where differences may be rooted in gender-based differences, though also provide representation for each of the three tiers of musical experience. Therefore, the potential influence of musical training can also be explored. Table 2.2 illustrates the distribution of participant demographics.

Gender	Musician	Some Music	Non-Musician
Female	3	3	3
Male	3	2	1

Table 2.2 The distribution of participants is designed to investigate two social dimensions (Gender and Musical Training). Both Gender and Musical Experience will serve as predictors in the analysis stage. Note: During exploratory analyses participants with Some Musical training and Non-Musicians were performing similarly and, as a result, were conflated into a single group.

2.7 Post-Experimental Data Prep and Extraction:

One way to explore the potential for acoustic-entrainment to changing pitch or intensity levels is to extract values at regular intervals from speech produced during test conditions, and then compare those values to others extracted from corresponding time points within speech produced in control conditions. *It is predicted that speech produced in test conditions will change in reliable ways to become more like their respective manipulations*. However, time points are less than ideal for comparison when considering duration/tempo changes. First, it is impossible to measure speech rate at a specific time point, as rate is a measure that must be calculated over some specified period. Second, changes to tempo result in time-expanded musical bars that, in turn, extend the run-time of the stimulus. Even if rate could be measured at specific time points (which it cannot), these time points would no longer time-align with Control conditions when considering the proportion of manipulation in a stimulus at a given moment. Therefore, extracting mean values for each musical bar, and/or splitting data from each trial based on whether or not the stimuli has been manipulated during the portion of speech production (that is, divisions by *Section*) may be more viable ways to approach the data.

Each of the above methods offers benefits and drawbacks, but may prove useful in the greater analysis. Extracting any of these values can be automated once textgrids have been generated for each recording, though creating textgrids by hand for each recording would be extremely time consuming. With some minor alteration and thoughtful scripting however, the textgrids described previously in the post-manipulation analyses can be altered to automatically extract the information of interest. Because these textgrids have been generated to accommodate both tempo-constant and tempo-altered stimuli, these templates have already been segmented by musical bar for all speech following the introduction of music. Though, background music is not introduced at the onset of a condition so this known segmentation begins well into each recording. The next step, then, was to recognize and segment any

lead-time (and speech) preceding the onset of music for each recording. This lead-time segmentation can then be concatenated with the appropriate pre-existing musical textgrid to gain complete segmentation for each trial/condition. The issue is made somewhat more challenging however, by the fact that the pre-music interval is somewhat variable in duration (~30 seconds) due to the manual execution of the experiment. Therefore, two primary Praat scripts were composed to automatically segment files and extract data while, importantly, maintaining synchrony with the musical bars as they exist in a given version of Science Music. The functions of these scripts are summarized in Appendix 2.

In short, pitch and intensity values were extracted from the recorded speech for each participant at every pitch pulse, including max, min, mean, and standard deviation values by Section (for each condition). *Section* was used to segment the stimuli on a gross level in addition to divisions by musical bar. As can be seen below in Figure 2.7, Section-intervals indicate whether there was (A) No music, (B) Music with no manipulation, (C1) Music with the onset of manipulation, (C2) Music where manipulation reaches the maxima, and (C3) Music where manipulation returns to origin values. Please note that Section, Bar, and Time-point are all temporal measures of different scope, providing related means of comparing the progress of a given manipulation through progressively finer grained levels of detail.

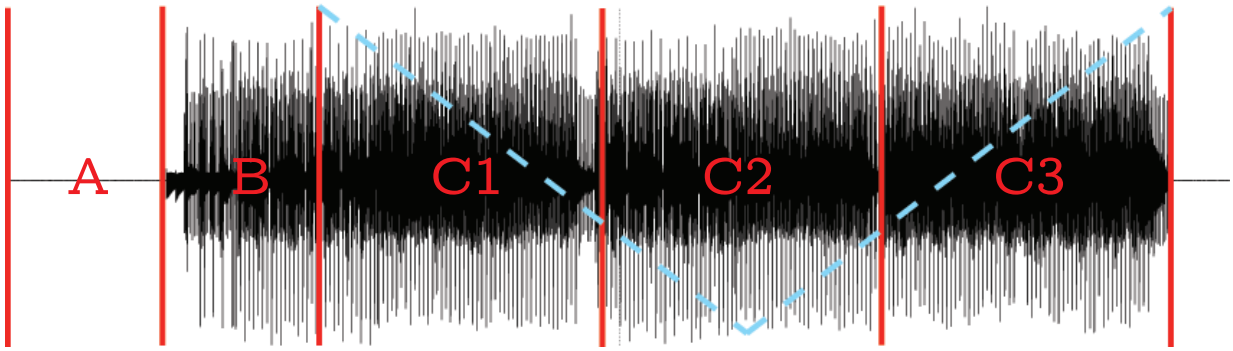

Figure 2.7 Within each Block (or Condition) the speaker's productions were split into 5 Sections (shown above) representing the degree of manipulation in the stimulus during portions of production. For example, Section A denotes speech produced before the introduction of background music, whereas Section B shows the introduction of background music before acoustic manipulation begins. Sections C1, C2, and C3 describe the onset of manipulation, the manipulation reaching its target/maxima, and the return to onset values, respectively.

Duration measures were extracted in a similar fashion – though on a coarser scale – again, based on wav/textgrid pairings. However, where intensity and pitch values are relatively straightforward to extract using pitch pulses and time points within the source signal, speech rate measures can *only* be extracted as averages over longer periods of time (e.g., most often as either syllables per second or syllables per minute, with the latter analogous to the standard musical beats per minute measure). Both of these rate measurements were calculated and extracted using LaBB-CAT (Fromont & Hay, 2008), an online platform for storing, organizing, and automatically annotating spoken corpora, in conjunction with the CELEX dictionary (Baayen et al., 1993).

2.8 Analysis

2.8.1 Overview and Exploratory-Analyses

The description following immediately summarizes numerous early attempts to understand data collected during the Pitch, Intensity, and Tempo-duration conditions. While these attempts informed the final analyses, the methods developed and eventually used are outlined in section 2.8.3.

At the onset of analysis it was unclear which statistical tests would be most appropriate due to the unavoidable autocorrelation issues which stem from time-series data. For example, regression models presume independence across observations and, barring some creativity during data-exploration, that assumption is violated when one applies the method to data collected over time at regular intervals. Importantly, when addressing autocorrelation in the data, Pinheiro & Bates (2000: pp. 226) explain that correlation structures can be used to model the dependence between observations. As a result, multiple analyses were explored before a suitable method to explain these data was selected after exploring random effects, autocorrelation terms, generalized additive models, and even the use of super computers.

First, using r (R Core team, 2013) as a platform for analysis, along with the rms (Harrell, 2016), LanguageR (Baayen, 2013), and lme4 (Bates et al., 2015) convenience packages, provisional linear mixed effects regression models were fit to explore the dataset using restricted cubic splines (the 'rcs' function). Linear regression models seemed a reasonable tool for such exploratory analyses given the fact that manipulations were imposed progressively and *linearly* toward a target value, before then returning to their original values in a similar linear fashion. If speakers are indeed converging with these acoustic manipulations, then use of this test is theoretically motivated when predicting linear changes over time in speakers' productions. This approach, however, does nothing to account for the autocorrelation in the data, which may be falsely inflating effect-sizes. Additionally, regression models presume linear (non-bending, unidirectional) effects, and effects within these data are predicted to change direction after a stimulus reaches its manipulation-target. To address this issue, *knots* (or, points where the trajectory of a linear effect can change direction while maintaining linearity on either side of that knot) were introduced through a restricted cubic spline to allow for aspects of speech to vary along with the stimulus. Four knots were included in all modeling, and the number of knots was selected based on predicted change given stimulus-manipulation i.e., one at the beginning and another at the end of a condition, plus one at the onset of manipulation, and a final knot was introduced for the change in direction after having hit the target value in the middle of the manipulated portion.

In these models either fundamental frequency (F0, measured in Hz) or Intensity (measured in dB) served as the dependent variable, depending on which was being modeled at the time; tempo/duration data were not yet available during this exploration. Here, relatively simple models were generated, exploring only the potential for a *Time * Condition* interaction. Specifically, these preliminary analyses were designed to investigate whether or not measurements taken from speech over time would change in predictable ways as a function of Condition (i.e., when one compares productions from within test-conditions to the control-conditions, though, note that all control conditions have been combined – that is, treated as alike – on the grounds of no predictable difference between

them). These analyses suggested convergence in the predicted direction while fitting both Pitch and Intensity models – however, not all assumptions of the tests had been met, and further reading on statistical methods brought to light the possibility of *additive modeling* as a more appropriate tool for exploring these data.

Generalized Additive Mixed Models, or *GAMMS,* do not presume independent observations, nor do they require all effects be linear in nature. In fact, one main reason for using this type of modeling is because it presumes the trajectory of effects will include at least some degree of *'wiggliness'*. Therefore, using the *mgcv* package (Wood, 2011), GAMMs were fit with structures much like the linear models described above (syntax within GAMMs is very much like the syntax of the linear mixed effects models, though the outputs and interpretations are rather different) exploring the potential for *Time * Condition* interactions. Indeed, because all statistical assumptions about the data had been met in this context, these tests were expanded to delve deeper into the dataset, investigating the potential for differences based on participants' *musical background, gender,* and other variables of potential interest.

Unfortunately, results were inconsistent based on the specifics of each model. Beyond the predictors motivated for inclusion (e.g., Time, Condition, etc.), there are three parameters that must be set appropriately when modeling with GAMMs (otherwise the default 'optimized' settings are imposed automatically). The settings of these parameters, being (1) The *type* of smoother used across observations, (2) The *amount* of smoothing allowed by that smoother, and finally (3) *The number of knots,* which, as mentioned above, specify the number of predicted bends (or the degree of 'wiggliness') in the data, can dramatically alter the fit of each model. In short, various combinations of settings could dramatically alter model outputs, and I was not in a position to use this tool as effectively as was needed. Given the mixed results observed through multiple models believed to be justifiable (resulting from different parametric settings), it seemed reasonable to once again explore the possibility of linear modeling, though some creativity was required to meet mathematical assumptions of the method.

Beyond issues rooted in autocorrelation, returning to linear modeling as the most appropriate tool for the job raises new problems due to the enormity of this data set (~1.5M observations in total, across all conditions). Specifically, when one includes terms to account for autocorrelation within the data, computations become much more complex, and the hardware requirements are also increased to make such calculations. With a data frame of this size, even simple models would often take as long as 6 days to fit on a personal computer. However, in response to this issue I explored the use of High Performance Computing (HPC) systems, often referred to as *super computers* or *clusters*. Unfortunately, at this time the University of Canterbury's onsite cluster had been decommissioned, and all HPC needs were redirected to the New Zealand eScience Infrastructure (NeSI) Pan cluster, housed by the Centre for eResearch at the University of Auckland (http://www.eresearch.auckland.ac.nz/en/centre-for-eresearch/research-services/computing-resources.html).[11] Using their system was a valuable lesson in how exactly super-computers work: Specifically, super computers are most effective when one can separate multiple tasks for processing in parallel to multiple nodes (conceptually: sub-computers). In this way, the computer can work on separate portions of a greater task or problem simultaneously. This type of task delegation, however, is for the most part impossible with regression modeling. In fact, recognizing that generating each model is a single task that largely disallows parallel

[11] Massive gratitude goes Francois Bissey and the University of Canterbury's High Performance Computing group for all the time spent helping me learn to use the cluster, as well as all the trouble shooting involved in trying to bend it to my will.

processing, the time-related issues mentioned above – which were the result of limited Random Access Memory (or, *RAM*) – became more troublesome using the cluster than running them on my personal desktop computer, which seemed to have more available RAM than the average node on the Pan cluster.[12] As a result, all models were run on my iMac desktop computer (2.7GHz i5 processor, 16GB RAM), taking nearly one week to complete every iteration of the modeling process.

Finally, as mentioned above, when using linear modeling to explore time series data one must consider how the autocorrelation is influencing model outputs. One approach controlling for these often inflated effects is to use an autocorrelation regression model which groups responses by participant over time, such as an AR(1) autoregression. This type of approach specifically incorporates the fact that a speaker's dependent variables (with respect to pitch, etc.) at time t are not independent from that same speaker's variables at time (t-1). It is, in theory, possible to include an autocorrelation term as well as random effects in a mixed model. However, in practice, models typically incorporate autocorrelation or random effects, but not both (see Pinheiro & Bates 2000; p. 254, 257). Presuming the fixed effects structures are identical across models, the random effects and autocorrelation terms can do much of the same heavy lifting, resulting in over-fitting the models when both are included. Therefore it is recommended that, having decided on a sensible fixed effect structure, one might then select the preferred random effects or correlation structure using ANOVA comparison of the otherwise identical models, and use the information criteria measures as a metric. Having explored the inclusion of both autocorrelation and random effects structures while modeling these data, the mixed effects models consistently outperformed the autocorrelation models. As a result, only mixed models will be described in section 2.8.3.

2.8.2 Predictor Variables and Varied-Approach Modeling in Three Steps

Information regarding the following variables was collected (in all experiments, to the extent applicable), and will be explored throughout the entirety of this dissertation:

- **Block** – Refers to where in the experiment a participant experienced a particular condition. For example, each participant experienced 5 conditions (i.e., (1) Test (2) Control (3) Test (4) Control (5) Test), and a value of "3" for Block would indicate that this was the third condition within a given session. (*A five-level factor*)
- **Passage** – Refers to which of the five passages a participant would read aloud during a condition. (*A five-level factor*)
- **Condition** – Refers to the acoustic treatment; namely, whether or not the stimuli heard within a block had been manipulated for Pitch, Intensity, Tempo/duration, or remained unaltered as a Control condition. (*A four-level factor*)

[12] There are two hi-memory nodes in the Pan cluster, but time with them is extremely limited and the wait-times to access them create further problems.

- **IdentGender** – Refers to a speaker's identified gender. Participants were asked what gender they identify with, and were given no prompting or leading toward binary divisions. (*Responses comprised a two-level factor*)
- **Musical Training** – Refers to the amount of formal musical training and/or time a participant spends practicing or playing an instrument regularly. Divisions in this category were made as follows: 1) Participants with no more than 6 months formal training over the course of their lives were considered non-musicians (NM); 2) Participants with no less than 6 years formal musical training AND/OR play and practice for a combined minimum of 10 hours a week were considered musicians (M); and 3) an 'other' category was created for all participants who did not meet the criteria outlined in (1) or (2), indicating the participant had some musical training (SM). Preliminary analyses suggested these SM participants performed very much like the NM participants, and these groups were eventually conflated as a result and the amalgam referred to as NM. (*A three-level factor*)
- **Age** – Refers to a participant's age in years. (*A continuous variable comprised of real integer values*)
- **Native Language** – Refers to a participant's first language (effectively, the language a participant spoke at home while growing up); this was eventually reduced to a binary variable indicating whether or not the participant was a native speaker of English. (*A two-level factor*)
- **Engagement** – Refers to how likely the participant was to listen to music in a style similar to Science Music, with the possible responses: (1) Yes, (2) No opinion, and (3) No. **(***A three-level factor***)**
- **HoursMusic** – Participants were asked to estimate, on average, roughly how many hours a day were spent listening to music. (*This was a continuous measure, comprised of real numeric values*)
- **Choose Music** – Refers to whether or not the participant would typically choose to listen to music while performing cognitively demanding tasks e.g., reading or writing academic papers. (*A two-level factor*)
- **Time** was segmented on three levels with the intention of including only one metric in a given model (with one exception necessitated by the Tempo condition); this choice was made to allow for different methods of statistical testing. Thus, (1) precise values were extracted at every pitch pulse to ensure all observations include measures for both F0 and intensity; (2) mean values were extracted by musical bar (i.e., every 4 beats within Science Music); and (3) mean values were calculated for each Section to reflect various points of structure regarding manipulation within the stimulus. Note that, where fundamental frequency and intensity can be extracted at precise time points, speech rate necessarily must be averaged over some interval. Because 1.8 seconds (roughly the duration of one musical bar in Science Music) is generally insufficient to extract a reliable speech rate, and because the duration of musical bar changes as a function of the Duration-manipulation, in the Tempo condition speech rate was extracted as syllables per second at regular 5 second intervals. (*Both 1 + 2 were continuous real numbers, whereas 3 was treated as a five-level factor*)
- **Proportion of Change (*PropChange*)** – Refers to the theoretical proportion of change within the stimulus at a given time point (theoretical only insofar as control conditions were not manipulated but could also be tested at specific time points). Values ranged from 0 (indicating no change) to 1 (indicating the

manipulation had reached the target value). This measure was included only when modeling data from sections C1-C3, and was calculated as follows (see equations (1) and (2) below): For each observation, the time point coinciding with the onset of manipulation was identified and referred to as 'Tstart' – this was the transition from Section B to Section C1 in each trial where, to avoid ambiguity, the first observation in SectionC1 was used and not the last cell in Section B. Each half of a manipulation – that is, the string of observations preceding and following the target value – was imposed linearly, where targets were reached at 88.5 seconds into each signal. Therefore, adding 88.5 seconds to Tstart recovers the point at which a manipulation reaches its target (referred to as 'Tmax'). We now have the start and end points of our first line. Values for each time point (*Ti*) in the line can then be used to extract a proportion-of-total-manipulation using equation (1):

(1)

$$\frac{Ti - Tstart}{Tmax - Tstart}$$

e.g., (Speaker 1)

$$\frac{Ti - 66.4808}{154.4708 - 66.4808 = 87.99}$$

While this methodology may seem overly complicated, it was necessary due to the experimental design. Because musical noise was introduced at roughly, but not exactly 30 seconds into a trial, using a single time value for Tmax in all trials would have been inaccurate. Furthermore, given both sampling frequency and the potential lack of pitch tracking for voiceless segments and pauses, there may not have been a Tmax sample falling at exactly the 88.5s mark of a trial (as in the above example). In such instances the nearest sample was selected as Tmax.

Much like the first half of the manipulation, retrieving proportion values for time points within the second half of the stimulus manipulation also requires both the start and end points of the second line. Tmax now serves as the starting point of line 2, and is already known. The end point of line 2 can be found as the final time-point of Section C3 (referred to as 'Tend'). Therefore, the equation to retrieve the proportion-of-total-manipulation for all points in the second line (Tj) looks as follows:

(2)

$$\frac{\text{Tend} - \text{Tj}}{\text{Tend} - \text{Tmax}} \quad \text{e.g., (Speaker 1)} \quad \frac{243.6308 - \text{Tj}}{154.4708 - 243.6308 = 89.16}$$

Please recall that the proportion of manipulation in line (1) is increasing while it decreases in line (2); this explains why the terms within the numerator have been reversed in the above equations. (*Continuous, numeric values between '0' and '1'*)

2.8.3 Analysis: Method

Using the variables described above as predictors, the following analyses explore the possibility of entrainment to each of the three acoustic dimensions of interest – Pitch (explored as variation in F0), Intensity (explored as variation in talker intensity), and Tempo (explored as variation in speech rate as syllables per second) – where each is described in three distinct sub-analyses:

1. In order to isolate potential effects driven by the introduction of background noise/music of a sufficient level, values (by Condition) have been compared from section A with those from section B, where section divisions serve as a coarse form of temporal segmentation (cf. Figure 2.7). Analysis of this type allows for comparison of productions in silence (section A) to speech produced while speakers hear un-manipulated background music (section B). Importantly, this method also provides baseline measures for speaker productions both with and without background noise. Because there is no preceding form of baseline, no sensible predictions can be made regarding how participants may respond to the introduction of background music; however, a main effect of *Section* would suggest there is *some* influence driven by the introduction of background music from silence.

2. With baselines in place, and because there is currently no precedent for entrainment in speech production to non-linguistic signals, it makes sense to next explore the entrainment hypothesis broadly by looking for trends in expected directions before presuming real-time compensations. Therefore, in the second analysis for each acoustic dimension, models compare values from productions in section B (music with no manipulation) to values from section C (combining C1-C3, where the entirety of the manipulation takes place). These analyses are meant to establish whether or not human speech changes in predictable ways when speakers encounter manipulated portions of background music. For example, where this Pitch manipulation lowers the tonal center of Science Music by 200 cents at maximum, it is predicted that speakers' F0 in the combined section C would generally be lower than in section B – that is, if speakers do

indeed converge acoustically with background music. A significant interaction of *Time* with *Condition* would suggest there is an influence of the treatment on speech production.

3. The third and final analysis described for each acoustic dimension looks more closely at potential entrainment using the proportion of manipulation to predict variation in speech over time. As mentioned above, measures of proportion are either theoretical or actual depending on the condition (that is, Control vs. Test respectively). This analysis is therefore restricted to only data from sections C1-C3 in each block, progressing over time. That is to say, for example, that as a manipulation progresses toward its target speakers are predicted to adopt corresponding acoustic changes within productions, following both the initial and secondary changes that bookend a manipulation reaching its target. A significant interaction of *PropChange with Condition* could suggest the measure may be predictive of responses to that treatment over time.

Recall that stimuli in sections A and B have not been altered in any way. As a result, any variation observed in speaker-productions *within* either of these sections is likely due to speaker prosody or some physiological motivation (e.g., differences in speaker intensity as a function of breath expenditure)[13]. However, as mentioned above, reliable differences *across these sections* likely reflect change related to the introduction of music in section B. Importantly, Lombard speech and surprisal-effects can be distinguished somewhat from certain entrainment-based effects following a comparison of sections A and B (where speakers have experienced no music in A and the introduction of un-manipulated music in B). Changes recognized in Section B (coming from Section A) *may* be attributable to the Lombard effect or a surprise/distraction from the introduction of background stimuli; however, following this comparison, Section B can then be used as a secondary baseline in comparison with Section C (the latter an amalgamation of C1 + C2 + C3, where all manipulation takes place). Put simply, any changes beyond those observed during the initial introduction of unaltered music should, in theory, be attributable to something other than the Lombard effect – namely, the manipulation treatments (with the potential exception of the Intensity condition, given the experimental design).

During analysis some of the predictor variables listed above were confirmed to be highly correlated e.g., ChooseMusic and Musical Training (participants who choose to listen to music during cognitively demanding tasks were often also those with formal musical training); Time and Section (which were, in fact, designed to be used individually as alternative forms of temporal segmentation in early data exploration), etc. Many of these predictors proved useful during early data exploration – for instance, finer grades of temporal segmentation allowed trends to be discovered before Section was adopted to address autocorrelation, and HoursMusic led to the discovery of Musicianship as a strong predictor. Their inclusion was problematic in later analyses though, due to (a) redundancy e.g., Time vs. musical Bar vs. Section, (b) insufficient data representation across factorial levels within the dataset

[13] It is possible that acoustic variation in speech following the introduction of music but preceding manipulated portions could potentially (even likely?) induce some forms convergence as well, though any such variation would be difficult to describe given the experimental design. Section B is therefore treated as a reference for comparison against section C.

due to low participant numbers, which could falsely inflate effects, or (c) correlation between variables, where another correlated predictor may better explain the data. The models described immediately below have therefore been reduced to their simplest theoretical forms (however, Block has also been included as a control variable in these models [14]) in order to present the most concise and reliable description of potential entrainment effects. More in-depth exploration of further predictors (including gender as a predictor) has been reserved for a replication of this study in Experiment 2. It should be noted, though, that the earlier exploratory analyses consistently suggested effects varied given participants' previous musical training. Therefore, given the predicted Condition by Time interaction, data were subset by Musicianship and analyzed separately in order to simplify interpretation of the expected three-way interaction.

As mentioned above, three linear mixed effects models will be described below for each test condition, sorted by manipulation. Each was fit in R (R Core Team, 2013) using the lme4 and languageR convenience packages (Bates, Maechler, Bolker, & Walker, 2015; Baayen, 2013). Random effects were included for Participant and Passage in all models, though random slopes were not included in this analysis due to low participant numbers (n = 15). Model structures are described below, and each sub-analysis was completed in the following three steps:

(1) Comparing section A vs. B (DV ~ Section + Block + (1|Participant) + (1|Passage)
(2) Comparing section B vs. C (DV ~ Section * Condition + Block + (1|Participant) + (1|Passage)
(3) Measures from section C only (DV ~ PropChange * Condition + Block + (1|Participant) + (1|Passage)

In most contexts sections C1-C3 were averaged and analyzed as a single section. In order to keep modeling computationally manageable, only observations from the treatments of interest and control conditions were included in each model (for example, in the Pitch analyses only data collected during Pitch and Control conditions were analyzed). 497,556 observations were extracted in total from speakers' productions during the Pitch treatment, 479,562 from the Intensity condition, and 1999 observations were extracted in total from the duration/speech rate condition (recall that for the Duration condition each observation was averaged over a 5-second chunk). Following Baayen (2008: pp. 73-76) all t-values exceeding +/- 2 were treated as significant. *Tables representing all model-outputs from each sub-analysis are available as Appendix 3*, where significant differences have been bolded for convenience.

[14] Block was observed to consistently improve model fit, but is in a sense inextricably confounded with Condition. Not only does Block describe the order in which participants encounter conditions, it also inherently encodes information about *which* conditions can exist in specific blocks due to the experimental design. That is to say, blocks 2 and 4 were *always control conditions*, and blocks 1, 3 and 5 were *always test conditions* – and once a given test condition had been encountered it would not be repeated in the same session. Thus, the effect of Block order could not be clearly identified due to limitations of the study, though it was sensible to include within these models given the inherent ties to Condition.

2.8.4 Analysis: Pitch (recall that in the Pitch condition musical pitch was first lowered and then raised)

2.8.4.1 Pitch A/B Comparison

When comparing Section A to Section B in the Pitch condition, 126,336 observations were collected in total. These data were subset for separate analyses based on musicianship however, leaving 51,285 observations for the musicians and 75,051 for the Non-Musicians. From silence, the introduction of music was associated with a drop in F0 for all speakers; though, the magnitude of this drop was greater for the musicians (Est. -1.8351, t = -6.523) than for the non-musicians (Est. -0.7406, t = -2.807).

2.8.4.2 Pitch B/C Comparison

Data collected during sections B and C from the Pitch and Control conditions included 431,565 observations that were similarly subset by musicianship for separate analyses (leaving 177,589 observations for the Musicians and 253,976 for the Non-Musicians). When comparing speech produced in un-manipulated music to that produced during the pitch-manipulation, specifically exploring the potential for a Section * Condition interaction, we find that musicians appear to invert the predicted entrainment effect (Est. 1.3488, t = 2.924) and there is no significant further change in the voice-pitch of non-musicians (Est. 0.46314, t = 1.092). These interactions have both been plotted below (Figure 2.8) with Section on the X-axis, frequency in Hz on the Y-axis, and different coloured lines indicating Condition. Both groups also show main effects of Condition (M: Est. -2.7942, t = -6.404; NM: Est. -1.85941, t = -4.694) and Block (M: Est. -0.5392, t = -3.278; NM: Est. 0.57100, t = 7.808), and a main effect of Section nearly reached significance for the Musicians (Est. 0.5297, t = 1.978).

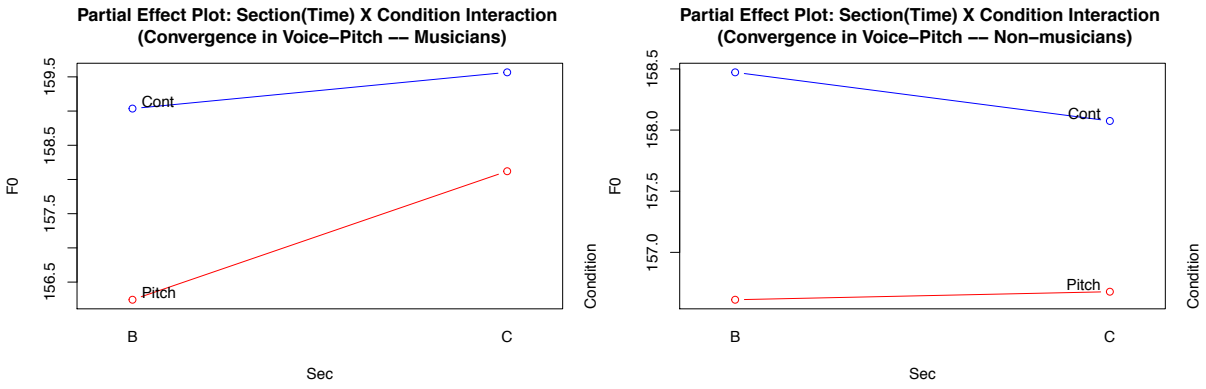

Figure 2.8 Partial effect plots illustrating the interaction between Section (a coarse measure of Time) with Condition. Musicians have been plotted on the left, and Non-musicians on the right. Recall that Section 'B' involved music with no manipulation, and 'C' was where signal-manipulations were imposed. Both plots express Section on the X-axis, frequency on the Y-axis, and use different coloured lines to indicate Condition. Note that we observe a significant effect of divergence on the left, where voice-pitch for musicians is rising as they experience the lowered pitch condition. Conversely, the non-musicians show no significant change in performance when the manipulated signals were introduced.

2.8.4.3 Pitch PropChange

Finally, 371,220 observations in total were included in the Proportion of Change Pitch-analyses (152,841 observations for the Musicians, 218,379 for the Non-Musicians). When including the PropChange measure to predict variation in speakers' fundamental frequency, musicians' performance appears to be reliably correlated with the amount of manipulation (Est. 4.4954, t = 7.513) – despite their inverting this effect – whereas the non-musicians' pitch trajectories cannot be predicted reliably using this measure (Est. 0.41192, t = 0.752). We also find a main effect of PropChange for Musicians (Est. -1.6924, t = -4.897), as well as main effects for Condition (M: Est. -3.6899, t = -10.283; NM: Est. -1.56635, t = -4.853) and Block (M: Est. -0.4453, t = -2.507; NM: Est. 0.26439, t = 3.361) for both groups.

2.8.5 Analysis: Intensity (recall that in the Intensity condition intensity was first raised and then lowered)

2.8.5.1 Intensity A/B Comparison

When comparing data from Section A to Section B in the Intensity condition, 120,947 observations were collected in total. Again, these observations were separated for individual analyses by musicianship (leaving 50,510 observations for the Musicians and 70,437 for the Non-Musicians). When music was introduced from silence, significant intensity-related increases were observed for both Musicians (Est. 0.99537, t = 26.22) and the Non-Musicians (Est. 0.66755, t = 21.84). Both groups also showed main effects of Block (M: Est. 0.16700, t = 8.78; NM: Est. 0.21033, t = 13.37).

2.8.5.2 Intensity B/C Comparison

Data collected from sections B and C during the Intensity and Control conditions included 416,689 observations; these were also subset for analyses by musicianship, leaving 176,291 observations for the Musicians and 240,398 for the Non-Musicians. Both groups exhibit significant interactions of Section by Condition, indicating nearly identical magnitudes when comparing intensity levels from speech produced during un-manipulated background music to speech produced during the intensity-based manipulation. In Figure 2.9 we see that speakers tend to be louder in the intensity condition, and that there is a general trend in both conditions to increase vocal-loudness over the course of a trial – however, the amount of increase observed during the intensity treatment is significantly higher than that observed in the control condition. Thus, data indicate that musicians are getting louder along with the manipulation (Est. 0.17656, t = 2.75), as are the non-musicians (Est. 0.187808, t = 3.57), who also showed rising intensity along with the manipulation. Both groups show main effects by Section (M: Est. 0.40241, t = 7.66; NM: Est. 0.612552, t = 13.97) and Block (M: Est. 0.06706, t = 6.08; NM: Est. -0.020114, t = -2.2). A main effect of Condition was also found for the Non-Musicians (Est. -0.828022, t = -16.71).

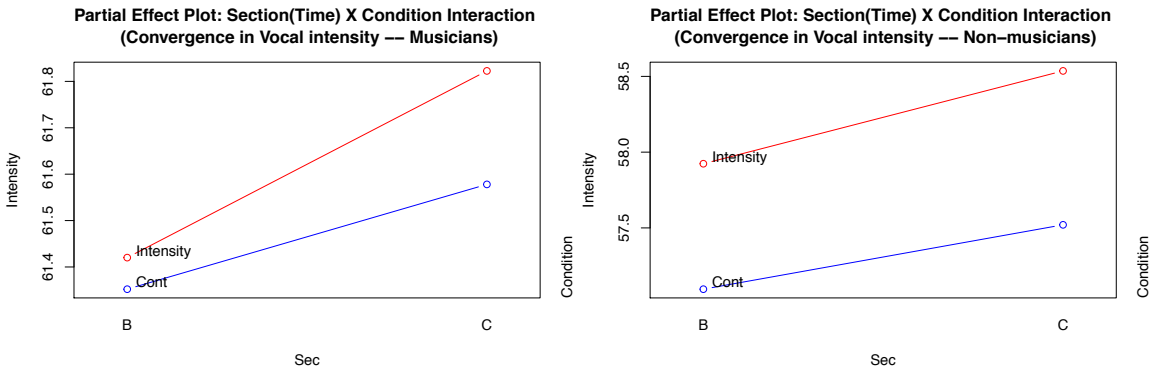

Figure 2.9 Partial effect plots illustrating the interaction between Section (a coarse measure of Time) with Condition. Musicians have been plotted on the left, and Non-musicians on the right. Again, Section 'B' involved music with no manipulation and 'C' denotes manipulated signals. Both plots express Section on the X-axis, frequency on the Y-axis, and use different coloured lines to indicate Condition. We observe significant effects of convergence on both the left and the right, where all speakers appear to increase vocal intensity along with increased intensity in the manipulation (over and above an apparent natural tendency to raise vocal-intensity in Section 'C').

2.8.5.3 Intensity PropChange

358,615 observations were included in the intensity-based PropChange analysis (151,818 for Musicians, 206,797 for Non-Musicians). In this case, when using the Proportion of Change measure to predict intensity variation we find Musicians' performance could once again be predicted reliably using this measure (Est. 0.22116, t = 2.341), whereas the performance of Non-Musicians could not (Est. 0.02795, t = 0.41). Main effects of Block (M: Est. 0.06594, t = 5.561; NM: Est. -0.040185, t = -4.125) and Condition (M: Est. -0.13428, t = -2.699; NM: Est. -0.995512, t = -24.753) were also found for both groups.

2.8.6 Analysis: Tempo (recall that Tempo was first decreased and then increased in this condition)

2.8.6.1 Tempo A/B Comparison

Data collected during Sections A and B during the Tempo/Duration manipulation included 487 observations; these were again subdivided for separate analyses based on the speaker's level of Musicianship, leaving a total of 199 observations for the musicians and 288 for the Non-Musicians. When comparing rate from productions in silence to rate from speech following the introduction of background music, we find no reliable change in rate for the Musicians (Est. -0.02705, t = -0.307), though we do observe a modest but significant decrease in rate for the Non-Musicians (Est. -0.19957, t = - -2.612).

2.8.6.2 Tempo B/C Comparison

Data restricted to those produced during sections B and C included 1,732 observations, which were also divided by degree of musicianship for separate analyses (leaving a total of 708 observations for the musicians, and 1,024 for the Non-Musicians). Comparing the rate of speech produced during un-altered music to sections where stimulus-tempo had been decreased, there is no significant effect of convergence for Musicians (Est. 0.257053, t = 1.725) or for Non-Musicians (Est. -0.01681, t = -0.128).

2.8.6.3 Tempo PropChange

Finally, 1,512 observations were included in the PropChange analysis for duration (that is, 618 observations for the Musicians and 894 for the Non-Musicians). Unsurprisingly, introducing the PropChange measure to predict variation in speaker rate revealed no effect for Musicians (Est. 0.10313, t = 0.574) or the Non-musicians (Est. 0.19360, t = 1.184) by Condition.

2.9 Discussion

In the previous analyses I investigated the potential for acoustic-convergence to background music in three acoustic dimensions, using a three-step analysis to broadly explore each acoustic dimension. First, the section A/B comparison explored values from section A (speech produced in silence) to values from section B (speech produced in unaltered background music) – these conditions were identical in every Block/treatment. Second, the B/C analyses compared observations extracted from speech produced in unaltered background music (section B) to speech produced while stimuli were manipulated in a systematic way (section C); on a coarse level, this analysis explores how manipulated acoustic signals may have influenced speech production. Finally, the PropChange analyses investigated the degree to which speakers/listeners' productions reflected characteristics of ambient noise on a fine-grained temporal scale; that is, these analyses explored variation in speech production as a function of the envelope of a given manipulation over the entirety of section C. A summary of the effects observed in Experiment 1 is available below as Table 2.3.

Acoustic Dimension (Manipulation)	A/B Comparison (no predictions)		B/C Comparison		PropChange	
	Mus	*NoMus*	*Mus*	*NoMus*	*Mus*	*NoMus*
Pitch (was decreased)	Y - decrease	Y - decrease	Y-inv.	N	Y-inv.	N
Intensity (was increased)	Y - increase	Y - increase	Y	Y	Y	N
Tempo (was decreased)	N	Y - decrease	N	N	N	N

Table 2.3 Summary of results for Experiment 1: "Y" indicates the presence of a significant Condition*Time (or Condition*Proportion Of Signal Manipulation) interaction, whereas "N" indicates a lack thereof. The abbreviation "inv." is used to identify situations where significant deviation was observed, though in a direction opposite to that predicted given the design of the test condition i.e., an inverted effect.

First, focusing primarily on effects observed during the Pitch condition, we see that fundamental frequency decreases for both musicians and non-musicians with the introduction of background music from silence. Other than the possibility of a typical Lombard response, there were no initial predictions for how speakers may behave after musical noise was introduced, as section B was in fact designed to serve as a baseline to which the manipulation portion of a condition could be compared as section C. However, participant responses were surprising in that the change observed for both musicians and non-musicians would not be predicted given previous Lombard research, as the Lombard effect is typically associated with rising F0 (e.g., Junqua, 1996; Cooke & Lu, 2010; Brumm & Zollinger, 2011) due to increased energy in the upper harmonics and increased tension in the vocal folds.

Considering the unexpected, but consistent decrease found for F0 along with the consistent rising intensity levels associated with the onset of background music for both musicians and non-musicians – the latter a change which *is* typically associated with the Lombard effect – it seems possible that participants may have honed in on some spectral component(s) within the signal (not unlike the idea of *home frequency* described above), and shifted their productions to become more like the stimulus in this dimension instead of exhibiting pitch-changes more consistent with a Lombard response. It is unlikely that this change in pitch would instead be related to a surprisal-effect, as Caelen-Haumont (2009) explains that surprise in speakers is also most typically associated with rising F0 (pp.97).

I have suggested that different speakers may converge to different aspects of the recording which, given the multiple voices which comprise the complex signal, draws attention to the fact that the theorized home frequency could exist above or below a speaker's unaffected voice-pitch. Though, considering where the majority of spectral information exists in Science Music, the home frequency should most often exist above a speaker's fundamental frequency. Work from Heffner & Heffner (2007) indicates the average human listener is not able to detect frequencies below ~60 Hz at presentation levels of 45 dB, suggesting that whatever spectral component speakers were moving toward might exist between that lower bound and each speaker's unaffected mean pitch – if this is, in fact, a convergence effect. Future work would benefit from devising a way to identify which specific components/voices within complex sounds speakers may 'latch on to' or find most salient, as this information would

both test the viability of *home frequency* as a concept, and also influence the interpretation of results of this and similar works to follow.

Interestingly, when global pitch is gradually lowered by 200 cents we discover the first change in behaviour that can be predicted by a speakers' previous experience. Having subset the data to allow for separate analyses that account for a participant's prior formal musical training, we find that the non-musicians show no further influence of the musical pitch manipulation. It makes sense then that their pitch trajectories in section C could not be predicted using the PropChange measure in light of this absent effect. The musicians' performance *could* be predicted reliably in this context using the PropChange measure; however, their pitch trajectories increased as a stimulus' global pitch decreased (and decreased when pitch of the stimuli increased i.e., an inverted realization of the predicted effect). It seems possible that musicians' previous experience talking against background music may drive these and similar changes in the name of intelligibility. If speakers are able to recognize and make use of spectral gaps through divergence, it seems such divergence could be useful in transmitting messages more effectively in the context of noise.

An arguably similar form of divergence has, in fact, been attested regularly in nature. Studies exploring behavioural plasticity and birdsong in a variety of avian species (primarily *oscines*, or, birds with learning-based song variability) have found that songs from birds living in urban areas "generally tend to vocalize at higher dominant frequencies" than do the same species living in rural areas, all other things being equal (e.g., Slabbekoorn, 2013; Nemeth & Brumm, 2010; Halfwerk & Slabbkoorn, 2009; Ríos-Chelén, Salaberria, Barbosa, Macías Garcia, & Gil, 2012). This stream of research argues the raised pitch observed in birdsong comes in response to the *anthropogenic noise* encountered in cities (low-frequency noise coming from motorized traffic, industrial areas and construction work, etc.). In order to make their songs better heard, these birds who diverge do so predominantly using two distinct strategies: Some birds actually sing the same songs in a relatively higher frequency range, while others choose to sing songs that include melodies with more relatively higher-frequency notes in them (and/or extend the durations of higher-frequency notes), thus raising mean pitch values in either context. It has been argued that this type of divergence in birdsong may even serve an evolutionary purpose (in addition to the communicative), where urban-birds who exhibit such plasticity may have a reproductive advantage over those who do not (Slabbekoorn & Ripmeester, 2008). Of course, these situations are not directly comparable, as the rural birds do not have to deal with the same environmental noise – though, these avian studies do suggest that some animals learn to raise vocal-pitch when regularly encountering certain types of environmental noise.

Therefore, revisiting the above-described rising pitch observed for musicians that comes with the lowered-pitch manipulation (i.e., the B/C comparison), perhaps we are seeing an effect not unlike that observed in avian research. If we can assume any degree of shared salience across speakers given previous experience, then perhaps the reason these speakers were observed to exhibit a rise in mean F0 is rooted in communication. Much like the birds whose drive is argued to be making their songs heard over the anthropogenic noise typical of 'city life', perhaps musicians in the present work are identifying a sort of spectral over-representation that comes with shifting music to concentrate energy in the lower end of the frequency spectrum, and therefore move productions collectively (and automatically) to avoid such competition and make their speech more intelligible for whoever/whatever might be

listening. If this is the case, then it makes sense that only musicians would show this effect due to a more acute ability to sense variation in the acoustic signals they encounter as a result of their training.

Though, importantly, if this is the case, then what we are seeing here is not so much an instance of acoustic convergence/divergence than some other compensatory mechanism in response to speech masking. Given the available information, there is no known way to distinguish between these possibilities. But it is also possible that speakers were in fact diverging from some resonant frequency (or frequencies) within the signals. Abstracting from work on speech, the convergence literature would suggest that if speakers were to find the background music disagreeable or annoying they might be prone to maximally distancing themselves from it acoustically (Babel, 2009). If the musicians happened to generally dislike ScienceMusic, and were more attuned to the manipulation given their highly developed listening skills, then this inverted effect (be it an effect of divergence or something else rooted in spectral competition) fits well enough with the known literature.

However, these lines of thought are not necessarily mutually exclusive. It is also possible that speakers are recognizing a spectral over-representation in the ambient noise (i.e., Science Music) and are driven by the notion of increased intelligibility (as opposed to finding the music disagreeable) to re-shape productions to become maximally different from the ambient noise. This possibility assumes that musical training results in a heightened ability to decompose complex acoustic signals, where spectral analysis in real time would allow speakers to compensate for various forms of noise by shifting speech to become maximally different from the masking/distracting signals. In this context, the effect would be more like typical forms of divergence, though with an experiential drive as opposed to a social one.

Recall that two competing hypotheses were described above regarding how participants with extensive musical training may respond to manipulated stimuli. The second hypothesis appears to be more correct thus far, at least with regard to pitch-based manipulations, inasmuch as the musicians showed differential effects from the non-musicians, and these pitch-based effects were in fact inverted. Of course, it is currently unknown whether this inversion was driven by divergence or other forms of spectral compensation; though, despite an inversion of the manipulated trajectory, the regularity of Musicians' behaviour observed in the PropChange analysis suggests they were in fact sensitive to acoustic variance that the non-musicians either could not recognize, utilize, or were not influenced by. At this point though, it is unclear whether the motivation for this inversion is social or experiential in drive (are musicians annoyed by the background noise, or compensating for it in the name of intelligibility?).

It should also be noted though, that the prediction regarding musical training was not wholly correct, as it stated musicians would likely be more *consciously* aware of these low-level acoustic changes. At the end of a session, as part of a debriefing survey, all participants were asked if they had noticed anything remarkable or different about the background music from condition to condition. Some discussion was allowed to ensure participants understood the aim of the question and, generally speaking, it was found that participants were unable to consciously recognize any acoustic manipulation in this task, nor changes to their speech, regardless of musical training. The fact that participants did not notice any of these changes was not completely unexpected; Natale (1975) remarks that in debriefing, no participants reported any awareness of any self regulation in a study exploring

convergence to intensity, supporting the somewhat automatic (or, non-intentional) nature of certain forms of entrainment.

Next, specifically exploring effects related to stimulus intensity, we find speaker-intensity levels rising (amongst other acoustic changes) for both musicians and non-musicians along with the introduction of background music from silence – this finding suggests acoustic change in the name of increased intelligibility. Recall that the known threshold for eliciting the Lombard effect is described by Lazarus (1986) as ~55 dB(A) for background speech and ~45 dB(A) for ambient noise. Because music is often described as sharing many characteristics with spoken language, and has even been noted for overlap with speech with regard to neuro-linguistic processing (e.g., Patel, 2008: pp. 9-238; Patel & Peretz, 1997: pp. 191-215; Koelsch, Kasper, Sammler, Schulze, Gunter & Friederici, 2004), one may question whether music is processed cognitively as more speech- or noise-like. While there is no conclusive evidence available through the present study to make such a distinction, it is noteworthy that the presentation of background music at ~45dB(A) was generally sufficient to elicit increased vocal-loudness. As a result, this finding may be interpreted as elicitation of a Lombard response through background music at this presentation level. Indeed, this finding may then be important for future works aiming to better understand and model the cognitive mechanisms associated with auditory processing, and how exactly human listeners process musical signals (that is, as relatively non-speech-like).

When the amplitude of a stimulus was gradually increased by 6 dB we found that speech from both groups of speakers exhibited further increases to mean intensity. However, once again it is unclear whether or not the effects observed in both of the above intensity-based analyses are an instance of convergence to an acoustic signal (at least to some degree, if only to maintain a constant buffer as a relatively consistent signal-to-noise ratio), or perhaps are nothing other than a Lombard response to environmental noise. Moreover, it is also unclear whether or not these processes are distinct in any meaningful way with regard to signal intensity. There is no strong evidence to suggest this initial rise in speaker loudness, nor the further changes observed along with the signal manipulation, are anything other than a typical Lombard response – though, such change can certainly be interpreted as a form of convergence to the background signal in that rising or decreasing intensity is typically matched to the envelope of the background noise. Simply, we are seeing that speakers' productions are reliably influenced by environmental noise in this way. Future works would benefit from devising a means to distinguish these potential drives for altered intensity, perhaps through a lowered-intensity manipulation that tests for entrainment below the known Lombard threshold.

In this context though, the effect size predicted through modeling was nearly identical for both groups, and this kind of uniform performance across speaker groups does support the change as generally indicative of a typical Lombard response; that is, more reflexive and less experience-driven. But again, the data suggest a speaker's previous knowledge and experience can influence the specifics of that response: Where the trajectory of a non-musician's changing intensity contour cannot be predicted through the PropChange measure, a significant PropChange * Condition interaction observed for the musicians indicates that their acoustic changes (or compensations) in the intensity condition reflected characteristics of the signal manipulation with a finer-grained degree of precision. Much like in the pitch condition, musicians appear to extract and use – or be influenced by – information within the signal in a way that non-musicians do not exhibit in their productions.

So, how then might we interpret these complementary results across PropChange analyses, which differ by musical group? Section 1.3 describes how musical passages and spoken language have both been shown to utilize loudness-related dynamics (i.e., relatively louder and quieter sections over time) and variation in pitch to important end. And, as no participants included in the current analyses noted any speech pathologies or exhibited any communicative deficiencies during a session, it seems very likely that all speakers should be (equally?) capable of extracting loudness- and pitch-related dynamics (i.e., speaker prosody) from the speech signals encountered in everyday contexts. Modeling these data suggests that the non-musicians may have been ill-equipped to do the same for background musical signals, at least on the same scale as the musicians. In other words, musicians' accumulated experience regularly deconstructing non-speech acoustic signals seems likely to have influenced the specifics of their responses to manipulations on a micro level.

Considered together, results from the analyses described in EXP.1 speak to the research questions stated above, regarding whether or not speakers' productions are reliably influenced by background music: (1) We have seen that participants' productions are altered, and altered in consistent ways (at least by group) when background music is presented at ~45 dB(A) during a speech-in-noise reading task. However, such effects were observed only during the pitch- and intensity-based conditions (admittedly, the present study may have not been sensitive to, and therefore missed, effects related to speech rate). Thus, while it is currently difficult to discern entrainment-based effects from other forms of noise-based compensations, the above analyses provide evidence that speakers' productions are influenced by ambient noise, and do not preclude the possibility of acoustic entrainment to background noise. Therefore, processes of convergence and divergence will be presumed below as possible until evidence is found to the contrary.

When considering the potential factors that might shape convergence/divergence (2) Evidence has been found to suggest potential social effects of divergence (seen through rising pitch, perhaps becoming maximally different from the background signal), or perhaps a more automatic compensation for a spectral over-representation. The contribution of a social motivation is yet unclear, though, given the role observed for musical training, an experiential drive appears to at least partially explain these effects. Both the intensity- and pitch-based analyses at least partially support a communicative impetus for these effects, as divergence from the pitch manipulation and loudness-matching to intensity would both result in increased intelligibility.

While there is no directly communicative context in this experiment, a potential limitation/confound of the study can be recognized in production materials which may have inadvertently influenced speakers: The fact that talkers reproduced prose *may* have felt more like a communicative event (i.e., storytelling) despite the removal of most punctuation (full stops were retained, but commas, quotations, exclamations, etc. had been removed). Impressionistically, it was noticed that readers tended to impose similar prosodic inflections while reading passages; if speakers felt like they were reading a story to someone, they may have been primed somehow to converge/diverge with background signals in a way that may reflect their attitudes about the music or readings. No information was collected regarding speaker-attitudes toward production stimuli, though these materials will be altered in EXP.2 to avoid the potential for such a confound. Conversely, information had been collected regarding participants' attitudes to the style of background music, where musicians were split fairly evenly across the three categories (like/no

opinion/dislike). It seems possible though that attitudes may have been somewhat less positive than reported, if the social drive was the root of pitch-divergence.

When exploring the potential roles of experience in these processes, the fact that musicians appear to be relatively more sensitive to fine-grained spectral- and intensity-based variation supports the finding that previous knowledge and experience can influence how speakers interact linguistically with their environment (cf. Hay, et al., 2017). It seems perfectly sensible that musicians' performance would be particularly susceptible to fine-grained acoustic variation, as reflected in the PropChange analyses. Because professional musicians have been trained to recognize low-level acoustic variation and to compensate for it in real time (or, even just to communicate effectively despite musical noise in the background), it seems possible the environment created through this experiment may have been treated and processed as at least tangentially similar. Indeed, acknowledging the lowered pitch manipulation was insufficient to further influence non-musicians beyond the wholesale effect observed through the introduction of background music further supports the claim from Hay et al. (2017) that certain compensatory mechanisms may not be available if a speaker lacks the appropriate experience.

As is often argued of the Lombard effect (e.g., van Summers, Pisoni, Bernacki, Pedlow & Stokes, 1988; Tam, 2017), it seems likely that the drive behind acoustic convergence and divergence may often be rooted in increased intelligibility. It further appears likely that the degree to which speakers are subject to various forms of acoustic convergence may vary by context, including differences in previous knowledge and experience which can serve to maximize speaker intelligibility or perhaps reflect speaker attitudes. Works reviewed above detail a variety of situations, linguistic and otherwise, where participants have converged with various aspects of their environment, and in a linguistic-specific context it seems very likely that convergence presents a means by which we aim to achieve more effective message transmission – be that transmission of the linguistic signal itself, or some underlying opinion/alliance. Whether speakers employ increasingly similar lexical items or syntactic structures confirmed to be known by their interlocutor (Brennan & Clark, 1996; Giles et al., 1991), or modify their vowel space to conform with that of their speech partner (Babel, 2012), a reasonable response to the "Why might speakers do this?" question can be provided through *effective message transmission* (e.g, Jenkins, 2000: pp. 167-175; Yazan, 2015; Currie Hall, Hume, Jaeger & Wedel, under review). Perhaps in the way of allotted resources it is statistically more cost effective to pre-emptively make small adjustments, progressively developing ways to share relatively more linguistic properties with the speech partner in order to avoid repetition and/or confusion. Extending this line of thought, we may aim to do the opposite with certain forms of background noise. For example, as suggested above, perhaps speakers diverge from the musical pitch in search of a spectral hole where they might land to compete less directly with the background noise. This theory could also explain why both musicians and non-musicians showed increased intensity during the relevant test condition and, in turn, why a significant PropChange interaction was only present in the speech of musicians; this was simply because *they are used to attending to such information* and using, compensating for, or dismissing it as required. However, it is still possible that inverted pitch trajectories of musicians may indicate a distaste for the background stimuli.

It should be noted that certain differences in effects may be the product of different numbers of musicians vs. non-musicians. While the above analyses seems relatively coherent, I draw attention to the fact that inequivalent

representation across these groups (as well as inequivalent representation in IdentGender) may have influenced outcomes in some important way(s). Equivalent participant representation is another factor, therefore, requiring increased efforts in the studies that follow-up on this experiment.

2.10 Conclusion

The present work has explored the potential for acoustic convergence to background music through a speech-in-noise reading experiment, testing specifically for convergence in the realms of voice-pitch, speaker intensity, and speech rate. Evidence has been provided to support the claim that speakers are influenced in reliable ways by ambient noise, and may be converging with or diverging from acoustic characteristics of background music in their speech. Moreover, a speaker's musical experience and training appear to influence whether we see reliably convergent or divergent characteristics in certain contexts. While it is unclear at this point whether or not participants are in fact converging/diverging or compensating for background noise in various ways to achieve effective communication, we do see systematic changes to speech production that warrant further exploration of this topic.

However, while the statistical methodologies adopted for analyses thus far have provided reasonable evidence, they were by no means without flaw – that is, other statistical tests may have been equally appropriate when analyzing these data. Indeed, issues involving participant numbers have resulted in an analysis with relatively low power, and results were largely unpredicted and interpreted after the fact. Therefore an experiment designed similarly, while addressing issues related to analysis and production stimuli would be of major benefit to supporting the current findings. Such a replication is available in the following chapter as EXP.2.

CHAPTER 3: Experiment 2 (Convergence to Background Music: Replication)

3.1 Introduction

While Experiment 1 (herein referred to as *EXP.1*) provides some potential support for acoustic convergence and divergence in speech to background music, it was not clear that convergence/divergence were necessarily the processes driving these altered productions. We saw that participants' speech was altered in reliable ways when speakers encountered certain background signals, though altered performance in the intensity condition could also potentially be explained by the Lombard effect (Lombard, 1911), and some participants' rising pitch could potentially be driven by a tendency to maximally avoid spectral over-representation (i.e., moving voice-pitch toward less dense spectral representation as opposed to actual divergence). Beyond these theoretical issues in interpretation, it must also be recognized that the study included some limitations in design and analysis. Consideration of these areas, and a largely post-hoc interpretation of results, brings to light specific changes which could be made in a replication study to allow for a less convoluted, more robust analysis. Moreover, participant numbers were low in EXP.1, which negatively affected the power of the analysis. Therefore, aiming to corroborate and better understand effects described in the previous study, the following replication was designed to incorporate more straightforward analyses and a task believed less likely to colour participants' speech production in unintentional ways. The goal of this replication was to maintain as much similarity as possible with the design of EXP.1, while providing data that avoid the autocorrelation and power issues encountered in the first study. Experiment 2 was thus altered and executed as follows.

3.2 Overview and Generation of Background Stimuli

As in EXP.1, Experiment 2 utilizes a speech-in-noise reading task to elicit speech from participants. The background stimuli presented as ambient noise were generated using the same base stimulus and delivery (i.e., a diotic presentation of Science Music, described in detail in the previous chapter) and methods very much like those outlined in EXP.1 (more on this below). Variants of the background music had previously been manipulated progressively over time, increasing – or decreasing, depending upon the manipulation – the value in that acoustic dimension toward a pre-selected target in a linear fashion. Such manipulation made sense from a theoretical perspective insofar as *proportions of manipulation* for each stimulus gradually changing over time could be highly correlated with variation in speech production, presuming convergence to non-linguistic signals does take place. However, the design presupposes speakers' productions will effectively change in real time, reflecting characteristics of the stimulus with a marginal lag. In fact, analyses of test conditions from EXP.1 suggest that some but *not all* participants adjust their productions in this way. With this variation in mind it seems possible the use of stimuli that

have been manipulated globally could then more directly identify effects of convergence. That is to say, instead of testing whether or not the proportion of manipulation within a condition reliably predicts changes in production over time within that condition, one might instead generate a baseline stimulus which maintains a relatively constant tonal centre and intensity level over time, and then compare speech produced during that condition to speech from two alternative stimuli-based conditions which maintain either a consistently lower tonal centre and unaltered intensity level, or lowered intensity level with an unaltered tonal centre. Presenting stimuli in this way has three major benefits: (1) It still allows for a general comparison of productions both within each condition and against the other conditions where, for example, one would expect that all productions from within a *lowered pitch* condition should generally result in speech with a *relatively decreased F0* when compared to the baseline productions; (2) Extracting mean acoustic values by item increases the number of independent observations, which avoids the methodological issues associated with analyzing time series data described in EXP.1 (re: autocorrelation); and (3) Statistical power is also increased through increased numbers of independent observations per treatment.

Background stimuli for EXP.2 were once again generated using Praat (Boersma & Weenink, 2014) and Ableton's Live 9 (Ableton, 2015). The Baseline/Control condition in this study is the same, un-altered version of Science Music experienced by participants in EXP.1; this stimulus provides a relatively constant amplitude and tonal centre, which can be compared in a straightforward way to the altered conditions by extracting mean values for a given acoustic dimension for each item produced by a speaker. The lowered Pitch condition was created once again using Live's *Warp* function by retracing the pitch envelope over time. However, where in EXP.1 manipulation involved multiple linear functions imposed over time to decrease and then increase the composition's tonal centre, in the present study the envelope remains flat but was set for the condition's entirety at 200 cents below the level in the Baseline condition. And where in EXP.1 the intensity manipulation was achieved by retracing the amplitude envelope over time, the Intensity manipulation for the present work was much less involved: The average intensity level of the un-altered sound file was assessed using Praat's *Get intensity (dB)* function, and the modified stimulus was then created by subtracting 6 dB from that value and using Praat's *Scale intensity…* function to decrease RMS intensity of the entire file by 6 dB SPL.

Note that Intensity was raised in EXP.1 while it is lowered in the current replication. This choice was made for three reasons: (1) Given the known threshold for eliciting Lombard speech through non-speech noise (Lazarus, 1986), this altered intensity manipulation seems a sensible way to potentially distinguish convergence to vocal intensity from the well-documented Lombard effect, provided the lower-bounds of intensity-related convergence extend below those of Lombard speech. In other words, if entrainment-based effects were observed at presentation levels below 45 dB, such behaviour would support convergence as distinct from the Lombard effect. (2) Another clear benefit of this altered manipulation is that it makes background stimuli more directly comparable across conditions – that is, they are all lowered in their respective acoustic dimensions. And, finally, (3) the Intensity condition was raised and then lowered in EXP.1 to explore the space between the known lower bounds for the elicitation of Lombard speech via non-speech- and speech-based noise (cf. Lazarus, 1986). Being unsure at the time of whether listeners might process musical background noise as more speech- or noise-like, and presuming a relation between potential convergence to intensity and the Lombard effect, it seemed reasonable to present the stimulus at

the lowest level known to elicit Lombard speech in order to gain some insight regarding how background music is processed cognitively.

Another important difference between EXPs 1+2 is the lack of a revised Tempo condition in Experiment 2; this choice was made for two reasons. (1) Analyses of the previous tempo data did not provide any indication of temporal convergence for Musicians or the Non-musicians, and (2) Given the various ways one might calculate speech rate as well as the many potential approaches to analysis which could be argued as theoretically motivated when analyzing speech rate, I have reserved replication of the Tempo condition for future study and it will not be mentioned below further. Therefore, three iterations of Science Music have been generated for use as background stimuli in Experiment 2: The original Science Music (SM) file serves as the Baseline/control condition; the lowered 'Intensity' condition, which was generated as SM – 6dB; and the lowered 'Pitch' condition which was generated as SM – 200 cents. Both manipulations observe the same manipulation target-distances selected for EXP.1 at maximum (though intensity is lowered by the same amount it was raised in EXP.1) to maintain relatively direct comparability.

All background stimuli generated for this experiment are available online for download through the following link: https://github.com/RyanPodlubny/ScienceMusic_EXP2

3.3 Controls and Considerations Regarding Experimental Design

Beyond an altered manipulation structure for background stimuli, there were also potential changes recognized to improve upon experimental design. Experiment 1 was structured so that each Session (S) for a participant was comprised of five blocks, where Test (T) conditions were separated by Control (C) conditions (i.e., S_i = T + C + T + C + T). In these sessions the presentation order of different Test conditions was randomized for each participant, and roughly two minutes of silence broke up all conditions to serve as a rest from the task. However, exploratory analysis of data gathered through EXP.1 suggested three confounds which could be addressed through simple design alterations:

a) CONFOUND 1: With a relatively fixed block order, exactly *when* a condition was encountered within a session was at least somewhat predictive of what that condition would be (i.e., all blocks 2 + 4 involve Control stimuli; all 1 + 3 + 5 would be one of the three test conditions). Importantly, the issue could not be explored fully in EXP.1 due to the lack of independence between Block and Condition – however, this issue is easily addressed in the current replication by counterbalancing all possible permutations of stimuli presentation order. If participants will encounter Baseline (B), lowered Intensity (I), and lowered Pitch (P) conditions within a session in this study, then the experiment to follow should aim for an equal number of participants encountering each possible presentation order of these conditions. In this new structure the position of the Control condition (B) will no longer be static and breaking up test conditions, but instead will cycle through all possible positions within permutations, just as the other test conditions do in this design. Therefore, all participants will experience one of the following sessions: BIP, BPI, IPB, IBP, PIB, or

PBI, where the position of a condition is no longer inherently predictive of what that condition may be. While this was not an issue for the statistics reported in EXP.1, it did prevent a more nuanced understanding of the data.

b) CONFOUND 2: Because stimuli were presented as described above in (a), there was never a *true* control condition in EXP.1. All participants had experienced *at least* one of the test conditions before encountering all Control conditions, which means that performance in these conditions may have been contaminated through previous experience in the session (i.e., the potential for persistent effects). The alteration described in (a) however, also places some control conditions (B) at the onset of a session, necessarily providing untainted data.

c) CONFOUND 3: Early provisional analyses of EXP.1 suggested the possibility of *cumulative effects* across test conditions inasmuch as the influence of a test condition may spill over into subsequent conditions (much like the description above in (b) acknowledging potentially tainted data). Data collected during Experiment 1 suggest it is possible, if not likely that two minutes of silence were insufficient to reset participants' auditory processes. There is currently no known agreed-upon methodology in the literature for resetting participants' cognitive and/or psychoacoustic mechanisms in between test treatments, though there are sensible methods that could be employed that are likely to be more effective than a brief silence. A new and unrelated – and intentionally differently distracting – task was therefore introduced to break up the three treatments in EXP.2. Therefore, in between each condition participants experienced a distracter-task in the form of a video game played in silence.[15] Given that distractor tasks have been shown to inhibit short term memory (e.g., Brown, 1958) and sensory memory (Gilmore, 1991; Murphy, Cain, Gilmore, & Skinner, 1991; Perkins & Cook, 1990 – cited in Herz, R. S., & Engen, T., 1996), I believe that having participants utilize other cognitive mechanisms, which are not (directly) linked to auditory processing, may relax potentially persistent effects through redirected attention and a lack of auditory input/stimulation.

3.4 Production Stimuli

A final issue was recognized in the production stimuli used in EXP.1, where participants read aloud connected passages selected from nature magazines. Despite nearly all punctuation being removed, speakers tended to adopt similar prosodic inflections as they progressed through a reading (e.g., similar phrase breaks, intonational patterning, etc.), possibly related to potential emotional investment or perhaps a communicative/social element implied by the reading material. This tendency can introduce problems as the degree to which a given speaker may alter/emphasize productions in certain ways will not be exactly the same as those observed in another speaker's productions and,

[15] To address the potential for cumulative effects, white noise presented at a constant dB level was originally considered for this task, however the possibility of this noise overloading/hyper-saturating listeners' sensory mechanisms seems equally possible in this context, and I have therefore opted to once again use silence.

more importantly, this type of influence is unintentional and indistinguishable from the influence of the background stimulus. Such changes are therefore unquantifiable. New production stimuli have been developed as a result, in the form of a phrase-list to limit prosodic variation across speakers through avoidance of prose. Instead of a few lengthy connected passages, speakers will now read from a list of many shorter items.

Production-stimuli in this experiment have been designed for draw without replacement from a single list of words and phrases ranging from 3 – 7 syllables. The primary drive for use of variable syllable lengths in the wordlist involves both comparability to EXP.1 and an aim to avoid influencing speakers in certain unintentional ways. The connected speech passages used in the first experiment were comprised of varying sentence structures and lengths with the intention of avoiding list intonations and other less natural production patterns that often result from repetitive tasks. Because production stimuli of uniform constructions could likely result in a similar type of unnatural speech patterning (Morrill, Dilley & McAuley, 2014), participants read aloud short phrases of varying syllable numbers and word frequencies.

Recognizing that pitch tracking algorithms work largely by assessing periodicity in voiced phones, production stimuli were designed to be comprised of only sonorant segments in an attempt to maximize the amount of data which could be analyzed for each item on the production list. Recall that Science Music maintains a run-time of 3:27. With this duration in mind, each production stimulus was allotted roughly 4 seconds of a condition; that is, stimuli were designed to require no longer than 3.5 seconds to read, process, and produce aloud, and were followed by an inter-stimulus interval of roughly half a second (the specific duration of each ISI was selected as a randomized value between 250 and 500 ms). Therefore, 162 unique items were generated to meet the criteria described above and cover the three conditions – that is, 54 items per treatment. A list of these items is available as Appendix 4.

Please note that using some lexical items of lower relative frequency is a necessary consequence of choosing words comprised only of sonorant phones as production stimuli. However, the potential for any unintended frequency-based effects has been controlled for through randomization of the stimuli set across treatments.

3.5 Equipment and Stimuli Calibration

In this experiment participants would alternate between sub-tasks through two side-by-side stations set up on a single desk. The first was required to run the experiment itself and to record participants' speech. Thus, EPrime 2.0 (Psychology Software Tools, 2012)[16] was installed on a Hewlett Packard EliteBook 850 G3 laptop computer, along with the necessary experiment files to visually prompt speakers and record their speech. The second involved the same late 2013 Macbook Pro (2.4 GHz Intel Core i5) described in EXP.1 – a *participant* account was created on this machine with only the video game "Quinn" (Härtel, 2009) installed.[17] Internet access had been disabled on this

[16] I owe a great debt of gratitude to Jonathan Wiltshire for his patience helping me implement this experiment in E-Prime.

[17] Quinn is freeware, very much like the more popular game "Tetris", where an assortment of different blocks are dropped into the screen one by one, and the game player is tasked with fitting them together without any gaps. When a full row of blocks are joined together as a line from one side of the screen to the other, that line will disappear and the player is awarded points.

account in order to avoid accidental distractions or participants being interrupted during the task. The sound had been disabled as well on this account in order to avoid participants potentially being influenced by the game's soundtrack.

Recording hardware was identical to that used in EXP.1: A Sound Devices USBPre 2 audio interface was used to capture speech in conjunction with a Beyerdynamic Opus 55.18 MK II head-mounted condenser microphone. Sessions were run in the same sound-attenuated booth described in EXP.1. All speech was recorded at 44.1 kHz and 16 bits. However, with regard to software, signals were routed through E-Prime in this study (and not Praat). EPrime had been programmed to generate a single .wav file recording all speech produced during each 3.5-second presentation (from stimuli onset). Participants heard background music and real-time feedback of their own productions through a pair of Audio-technica ATH-M40x closed back, isolation headphones; the entire setup can be seen in Figure 3.1. In order to ensure all participants experienced background music at the predetermined presentation levels i.e., ~45dB(A) for non-intensity-manipulated conditions, equipment was calibrated using methods identical to those described in EXP.1. Photos illustrating the calibration setup are available as Figure 3.2.[18]

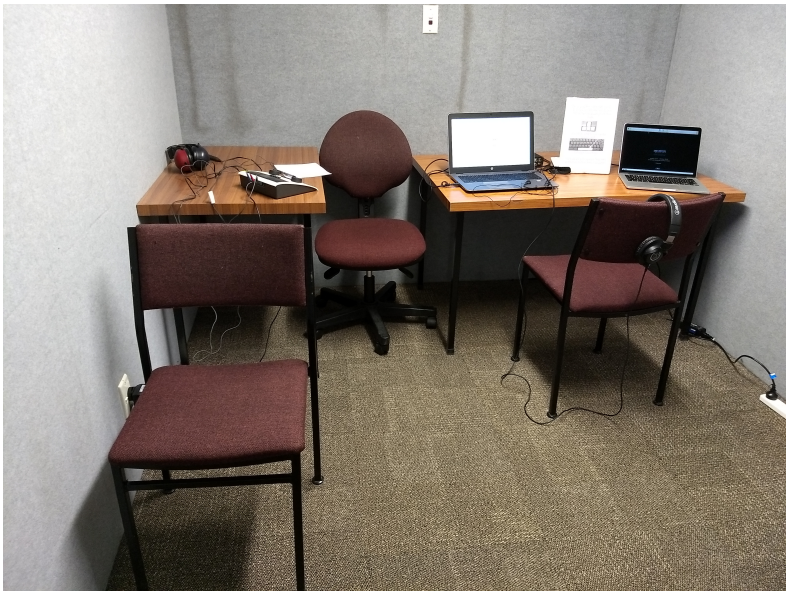

Figure 3.1 The experimental equipment as it was set up for each session in the sound attenuated booth. From bottom left in clockwise fashion: A chair for participants to face away from the experimenter during the hearing screening; the hearing screening station with headphone and the Interacoustics AS608 audiometer on a table; a chair for the experimenter during the screening; the HP Elitebook running the experiment; the USBPre audio interface; a sign instructing participants how to navigate the two experimental platforms; the Macbook Pro running 'Quinn'; a chair for the participants during the experiment; and the ATH40x headphones resting on the chairback.

[18] Many thanks are due to Julian Phillips for his help in calibrating the equipment for this study.

Figure 3.2 The same Brüel & Kjær model 4100 head and torso simulator used for calibration in EXP.1 is shown 'wearing' the Audio-technica ATH40x headphones used by all participants in this study. On the left we see the experimental setup as participant would later experience it; on the right we see a close-up of the ATH40x headphones during calibration.

3.6 Participants

As specific differences in performance between native and non-native speakers cannot be predicted at this point, and to maximize comparability with EXP.1, only native speakers of New Zealand English participated in the present study (n = 32). All speakers were recruited via posters displayed on campus, adverts through various forms of social media, and brief discussions in entry-level linguistics and music courses. Participants received a $15 NZD voucher for use at a local shopping centre in exchange for their time.

Based on the findings of EXP.1, formal musical training is expected to be an important predictor. Professional musicians have been trained to recognize subtle acoustic variation, both consciously and unconsciously, and research has shown that musically skilled subjects often perform differently than unskilled participants in rhythm-based tasks and certain neurological responses (Duke, 1994; Drake, 1998; London, 2012: pp.172; Chen et al., 2008). EXP.1 provides support for such effects. In the present work participants were again recognized as belonging to one of three groups: (1) Non-musician (< 6 months formal training in life), (2) Musicians (> 6 years formal training in life, AND/OR currently averaging no less than 10 hours per week performance/practice time); or (3) Some Musical experience (this is effectively an 'else' category which includes participants who fall in between the two more extreme categories). Once again, to the extent that it was possible, efforts were made to counterbalance for musical training and for identified gender, and the breakdown of participants was as follows: 16 had some musical training (SM), 8 were classified as musicians (M), and 8 had no musical training (N); 21 of these participants identified as female, and 11 identified as male. No participants reported identifying as non-binary.

3.7 Procedure/Protocol

Upon arrival subjects were assigned a participant number encoding meta-information formatted exactly as in EXP.1 (i.e., the experiment name, the year of participation, the participant's chronological rank in the study, identified gender, and level of musical training). Subjects were asked to read through and complete the same consent form and language background survey used in the initial study before the experiment began. Again considering a hearing screening more reliable than self-reporting, participants sat a full hearing assessment before taking part in the experiment; the testing was as described in EXP.1, though in EXP.2 an Interacoustics AS608 screening audiometer was used. All participants whose screening results showed any cause for concern were referred to the University of Canterbury's Audiology department for further assessment. These subjects were also excluded from the remainder of the study and all later analyses (n = 3, over and above the 32 participants who provided usable data).

All participants whose hearing fell within the acceptable ranges next sat the experiment. These subjects were outfitted with the equipment described above and seated at the experimental station within the sound booth. As mentioned previously, participants encountered three counterbalanced test conditions (Baseline, Intensity, Pitch) in a session, where treatments were broken up by an unrelated task (i.e., the video game). In each test condition production stimuli were presented visually on a computer monitor for participants to speak aloud – each production stimulus remained onscreen for 3.5 seconds, and was replaced by a blank screen before presentation of the following stimulus. The duration of the blank screen was a randomized value between 250 and 500 ms to avoid inadvertently falling into a regular cadence as a result of regular spacing during list productions. The presentation of production stimuli continued in this way for the entire run-time of each iteration of Science Music. With the exception of the lowered-intensity condition, all stimuli were presented at ~45 dB(A) as in EXP.1. In between each Test condition, participants were instructed to shift to the adjacent station/laptop in order to complete one level in the video game in silence (Figure 3.3, right). In this context one level of "Quinn" involved playing either (1) until the participant had scored 10 lines or (2) unintentionally ended their turn by stacking game pieces to the top of the screen. Signage had been placed in between the two platforms as part of the experimental station to remind participants how to navigate the different conditions (see Figure 3.3, left). Following the third treatment, participants completed the same debriefing survey used in EXP.1.

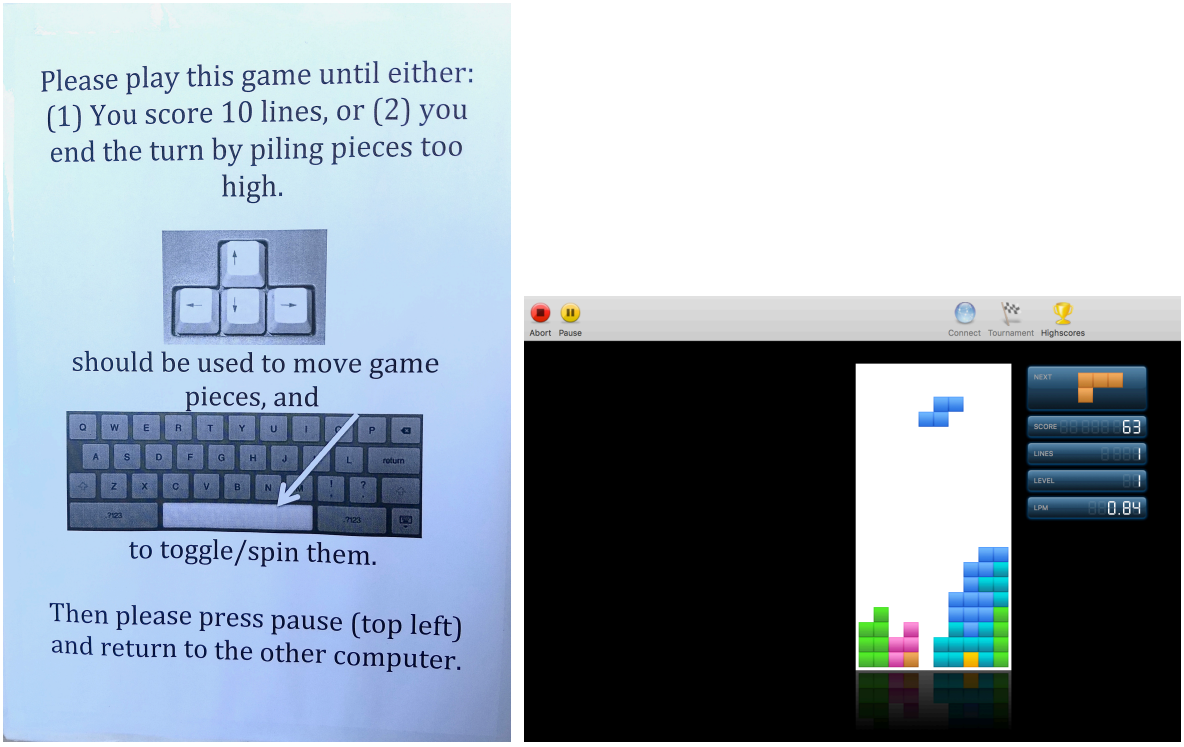

Figure 3.3 (Left) ***The sign available to participants instructing them when and how to work between the different experimental tasks; (right) "Quinn" as the participant would see the screen.***

3.8 Analysis: EXP.1 Replication

The following analyses were conducted in R (R Core Team, 2013) using the lme4, languageR, and lattice convenience packages (Bates, Maechler, Bolker, & Walker, 2015; Baayen, 2013; Sarkar, 2017). Unlike EXP.1, which required multiple comparisons within each condition to explore the influence of the test stimuli, the current (altered) design requires only one model per condition to explore such effects because the Time variable is no longer relevant. However, two approaches to analysis were employed to explore the current data: (1) The simple models fit to comparable conditions in EXP.1 have been applied to the present data, where the outputs are reported and contrasted to previous findings. Importantly, with time no longer a variable in this study model-predictions involve only a Musician * Condition interaction with inclusion of Block as a control variable; though, random intercepts for Participant and Item were also included when testing these data for the effects observed in EXP.1. (2) A series of post-hoc analyses were also run and will be described in detail in the section below entitled *Analysis 2* (Section 3.10).

Please note: Preliminary analyses indicated the sub-group of speakers with Some Musical training (the *SM*s) were performing differently than the Non-Musicians (*the Ns*) in the present study, where these two groups had been consolidated in EXP.1 on grounds of similar performance. Because of the relatively increased number of participants in this experiment, which allow statistically for the tolerance of more groups without deflating the power of the

analyses, these groups have been treated as distinct. Therefore, this difference in performance across studies has not been treated as cause for concern.

Based on the findings of Experiment 1, Condition * Musician interactions would be predicted. Specifically, testing the same models trained during EXP.1 should find effects where Musicians diverge in voice-pitch and non-musicians show no significant difference in voice-pitch. No predictions can be made for the SMs. Conversely, with regard to the intensity-based treatment, it would be expected that both groups would converge with the lowered signal intensity – that is, so long as convergence thresholds extend below those of the Lombard effect. If the lower limit for eliciting these two effects is shared across processes, then it would be expected that no significant change should be observed for either the musicians or non-musicians. Once again, no informed predictions can be made for the SMs.

5,184 observations were collected across the three conditions, each containing mean pitch and intensity for a given spoken item.

3.8.1 Replication: PITCH

Data were reduced to include only observations collected during the Control and Pitch conditions, as in EXP.1. Following this reduction 3,456 observations remained. Musicianship, Condition, and a Musician * Condition interaction were tested with Block included as a control variable. The model did not contain any random slopes, though random intercepts for Participant and Item were included. The optimizer was set to "bobyqa". When fitting the model from EXP.1 to data collected from participants in EXP.2 we find that both Non-musicians (Est. 2.5702, t = 0.145) and participants with Some Musical training (Est. 28.5090, t = 1.392) do not appear to be performing differently than the Musicians (i.e., no main effect). Releveling with SMs as the intercept shows that Non-musicians and SMs are also not performing differently than each other in this treatment (Est. 25.9388, t = 1.462). While trending in the predicted direction, the lowered mean fundamental frequency observed in the Pitch condition did not reach significance when testing for a main effect of Condition (Est. -1.5888, t = -1.607). However, when exploring the Musician by Condition interaction, plotted below as Figure 3.4, we see that non-musicians exhibit significantly higher mean fundamental frequency in the Pitch Condition than do the musicians (Est. 4.6659, t = 3.328), where this difference did not reach significance for the SMs (Est. 2.1665, t = 1.764). Releveling with SMs as the intercept shows that this difference was also significant when comparing performance of the Ns and the SMs (Est. 2.4994, t = 2.013).

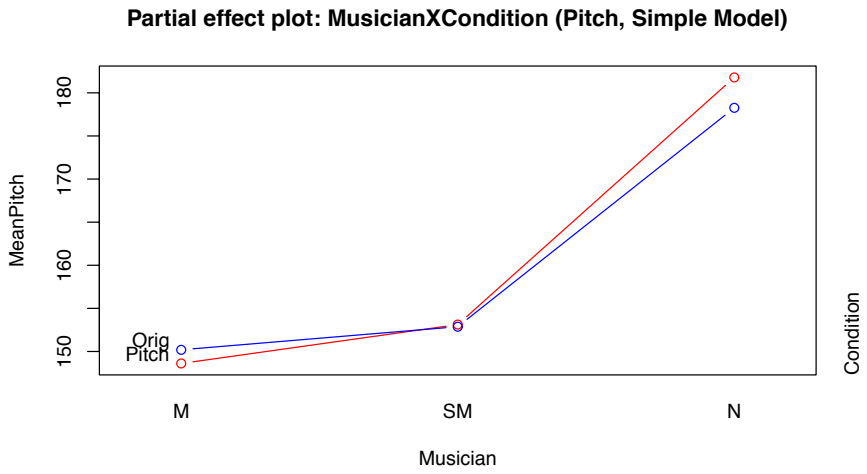

Figure 3.4 A partial effect plot visualizing the Condition by Musician interaction observed in the Pitch condition while extending the simple models from EXP.1. In this figure we see degree of musicianship on the X-axis, pitch frequency in Hz on the Y-axis, and different coloured lines to indicate condition.

While we see that these groups are at times performing differently from each other, it makes sense to next subset the data by musicianship to test for by-Condition effects for each sub-group; this within-group test will shed light on whether or not musical experience is in fact resulting in differential performance across conditions. In other words, sub-setting in this way controls for the difference between altered-performance driven by personal experience (i.e., through musical training) as opposed to the influence of the treatment, while still accounting for that personal experience. Please note that the baseline in these models was set to the Original condition. Reducing the data to observations collected from Musicians left 864 data points. Though voice-pitch appears to move in the same direction as the manipulation, in this context the within-group by-Condition difference is not significant (Est. -0.738, t = -0.673). Subsetting the data to test for by-Condition differences for the SMs only left 1728 observations; this difference also did not reach significance (Est. 1.0162, t = 1.684). Finally, sub-setting the data to test for a by-Condition effect for only the Non-musicians left 864 observations, where this difference was again not significant (Est. 1.987, t = 1.553). Thus, it appears the trends are significantly different from each other, but are not significant themselves when considered in isolation.

3.8.2 Replication: INTENSITY

Much like the preceding Pitch-analysis, data were reduced to include only observations collected during the Control and Intensity conditions – thus, 3,456 observations remained. Musicianship, Condition, and a Musician * Condition interaction were tested with Block included as a control variable. The model did not contain random slopes (as in EXP.1), though random intercepts for Participant and Item were included. The optimizer was set to "bobyqa". When testing the model trained in EXP.1 on data collected from participants in EXP.2, again, we find no main effect for Musician, indicating that neither the Non-musicians (Est. -2.9909, t = -1.67) nor participants with Some Musical

training (Est. -1.329, t = -0.88) were performing differently from the Musicians. Releveling with SMs as the intercept shows that Non-musicians and SMs are also not performing differently from each other in this condition (Est. -1.59195, t = -1.05). Though trending in the predicted direction, we find no main effect for Condition (Est. 0.158, t = 1.39), suggesting that there is no statistically significant difference in mean speaker intensity across treatments. Testing for a Musician by Condition interaction indicates that across conditions the Non-musicians are not performing significantly differently from the Musicians (Est. 0.2434, t = 1.57); however, the participants with Some Musical training do appear to perform differently than the Musicians in the lowered pitch condition (the partial effect is plotted below as Figure 3.5). Releveling with SMs as the reference indicates that these participants are also performing significantly differently than the Non-Musicians in the test condition (Est. 0.72164, t = 5.05).

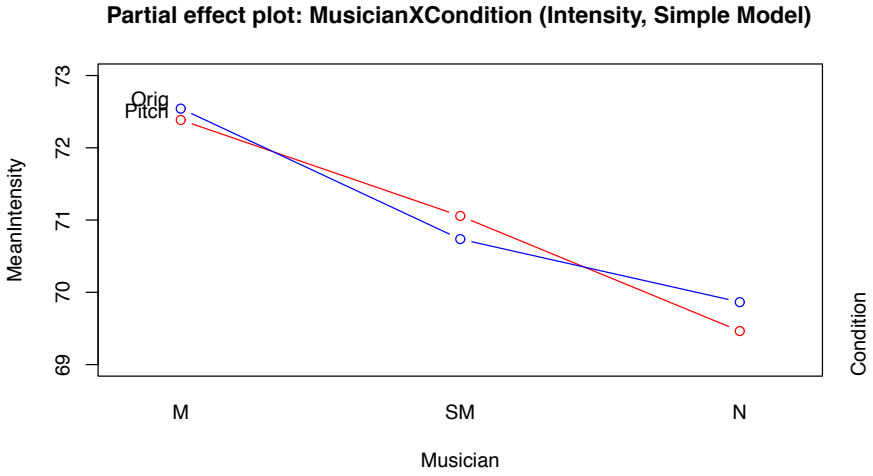

Figure 3.5 A partial effect plot visualizing the Condition by Musician interaction present in the Intensity condition while extending the simple models from EXP.1.

As above, it makes sense to next subset the data by participants' degree of musicianship in order to test the validity of any by-Condition effects that may be driven by degree of musical experience; please note that the lowered-intensity condition currently serves as the baseline in these models. Reducing the data to observations collected from Musicians only left 864 data points. These simple models suggest Musicians are not performing differently between the two conditions (Est. 0.027605, t = 0.20). Sub-setting similarly to test by-Condition differences for the SMs left 1,728 observations, and this difference also did not reach significance (Est. -0.14330, t = -1.62). Finally, restricting the data to test for a by-Condition difference in performance for only the Non-musicians left 864 observations, where the Non-musicians appear to be significantly quieter in the lowered intensity condition than in the unaltered (relatively louder) Baseline condition (Est. 0.7163, t = 6.23) – that is, exhibiting convergent behaviour.

3.9 Interim Discussion: Replication Analyses

Effects observed in Experiment 1 were largely not corroborated when comparable models were fit to data collected during EXP.2. In fact, much of what we find does not quite agree when comparing model-fits across the two studies. Where in Experiment 1 Musicians were diverging in the pitch condition, in EXP.2 the Musicians showed no significant pitch-based effect (but trended toward convergence); the Non-musicians, on the other hand, exhibited no significant pitch-based effects in either study. With regard to Intensity-based treatments, both the Musicians and Non-musicians showed effects of convergence in EXP.1, whereas only significant effects of convergence were observed for the Non-musicians in EXP.2. No significant intensity-based effects were observed for musicians in either of these analyses. These models suggest the Non-musicians of EXP.2 are performing more like the Musicians of EXP.1, and vice versa. Thus, predictions regarding this analysis were not met.

While it is possible that differences in the production stimuli (i.e., Connected passages in EXP.1 vs. Word lists in EXP.2) and in stimuli manipulation (i.e., Gradual manipulation within conditions in EXP.1 vs. Global manipulation across conditions in EXP.2) may have contributed to differences in speaker performance across studies, this seems a less likely explanation than either imperfect methodology in the first study, or a low-powered analysis due to too few participants. Statistical power was relatively low in EXP.1 given the methods used to deal with the autocorrelation issues, and increased participant numbers in EXP.2 suggest these recent data should provide a more robust analysis.

One important note involves *the direction* of manipulation in the current Intensity treatment. Intensity levels were gradually raised in EXP.1, whereas they were lowered to a point below the known threshold for eliciting the Lombard effect in EXP.2. If these results hold, where non-musical speakers converge below this threshold, then they may provide some evidence for entrainment to intensity being a process distinct from the Lombard reflex. If this is the case, then perhaps Musicians were less inclined to lower vocal loudness due to experience communicating in noisy environments, and would normally want to ensure they were heard above competing music in the speech environment.

Though, because there were multiple approaches to analyzing data in EXP.1, which were all theoretically justifiable but resulted in incongruent outputs when contrasted, it would be beneficial here to entertain the possibility of alternative methods where statistical outputs hold across complementary tests. Moreover, it seems likely those models may not have been optimized to the present data given the relatively sparse dataset they were trained to in EXP.1. One next step, then, would be to re-examine data from EXP.2 with an amended analytical approach, one that allows for maximal retention of the raw data as well as robust statistical power. Such post-hoc analyses are described directly below.

3.10 Analysis 2 (Post Hoc)

Re-analyses of the data collected during EXP.2 will be broken into two overarching sections, exploring two important questions: (1) *Is the variation observed across speakers and productions random?* Until now it has been

assumed that a given participant would respond similarly (i.e., converge or diverge) when experiencing various ambient acoustic treatments throughout the present work. Specifically, in EXP.1 it appeared that some participants converged, and others diverged. In light of the above analysis, it is unclear whether or not musical training is in fact a viable means to sort those who exhibit convergent patterns from those who diverge. Indeed, it is not even clear at this point whether speakers generally tend to converge or diverge, or might do one and then the other in various contexts. No longer presuming the former, and aiming specifically to test the regularity with which speakers tend to converge/diverge, the following analyses test if speakers exhibit significantly different behaviours in treatments vs. controls, and whether or not behaviour in one treatment is indicative of responses in the other. While it is possible that speakers are being influenced by these acoustic manipulations, analyses thus far do not discount the possibility that speakers may in fact not be influenced by these treatments, and the "effects" observed are instead noise from a random distribution of potential behaviours.

Therefore, in section 3.10.1, this assumption is tested through use of data transformation, descriptive analyses, and random forests while exploring whether or not differences and similarities in performance across conditions appears to be systematic for participants, or if there is no apparent link between potentially altered productions in the two conditions. Also, (2) *If performance across conditions does not appear random, then which variables might best explain what participants are doing?* In section 3.10.2, any variables observed to explain the data reliably through random forests and variable importance plots will be used to feed linear mixed effects models. These models will allow for a better understanding of how predictor variables contribute to any such effects through the explicit inclusion of interactions in the modeling process.

3.10.1 Is the Variation Observed Across Speakers Random?

Reanalysis first involved a descriptive approach to the data in order to explore specific assumptions regarding speaker behaviour. While modeling from EXP.1 suggested musical experience was driving effects of convergence and divergence, the analyses described directly above resulted in somewhat conflicting model outputs. While effects of convergence and divergence are observed, in Experiment 2 it seems not simply a case that musicians diverge and non-musicians converge as in EXP.1, nor is it necessarily the case that a single participant would even exhibit similar convergent- or divergent-behaviours across treatments. The plot, available below as Figure 3.6, shows t-values generated by participant for both test-conditions. These values were calculated using the aggregate function (and unpaired t-tests) in R to generate normalized distance measures expressing both effect-size and direction by condition for each participant by comparing observations in each test condition to corresponding observations from the control conditions (e.g., mean F0 from items in the lowered pitch condition was compared to mean F0 from items in Control). Therefore, each value is effectively a distance-score comparing performance in a given condition to that in the baseline.

As a brief point of note, I draw attention to the fact that paired t-tests may have been more appropriate in this context if the tests were primarily being used for inferential analyses, as the test compares samples of equal sizes (same number of observations across conditions) which were produced by the same subject. However, the function

serves here primarily as a data-transformation, and all data have therefore been subject to the same transformation. By this reasoning use of the test in this way is not an issue, and at worst results in more conservative t-values for these data. Unpaired tests, however, are more appropriate for analyses described later in this work (chapter 5) when transforming observations from samples of unequal sizes. Because data from the three experiments are eventually combined in a unified analysis in Chapter 5, I have opted to use unpaired tests for all t-value transformations in order to retain direct comparability for data collected across studies in that analysis. The t-value transformation results in one value for each participant for each condition – as a result, data from EXP.2 have been reduced to 32 pitch-based observations, and 32 intensity-based observations for use in the following descriptive analysis and with the random forests (also described below).

One can see in Figure 3.6 that, while certain participants show generally lower t-values in treatments than in the baseline conditions (that is, in line with manipulations), performance across conditions is irregular with regard to positive vs. negative effects. In fact, some participants are found to converge in one condition while diverging in another. These results might be expected given previous work in acoustic-phonetic convergence to speech, where Pardo (2013) explains in a review article that:

> "Talkers can converge in one dimension at the same time that they diverge or produce random variation in other dimensions. Moreover, this flexibility extends across different items. For example, convergence in F0 on one item does not imply that the talker will always or only converge on F0. On another item, the talker might converge on vowel formants or duration instead. Indeed, this is the kind of pattern that has been found in studies that have examined multiple acoustic attributes."

Therefore, if one thing is clear from previous work, it is the fact that *acoustic-phonetic convergence is often irregular*. The degree of synchronization in speech is well known to vary from one speaker to another, from one acoustic dimension to another, and even from item to item – at least, this is known to be the case when studying talkers converging to speech partners. As a result, convergence in speech production cannot be expected to constantly shape a speaker's productions, or at least not to always shape them consistently; this knowledge helps explain the variation observed below in Figure 3.6, which shows t-values (on the X-axis) presented for each participant (on the Y-axis).

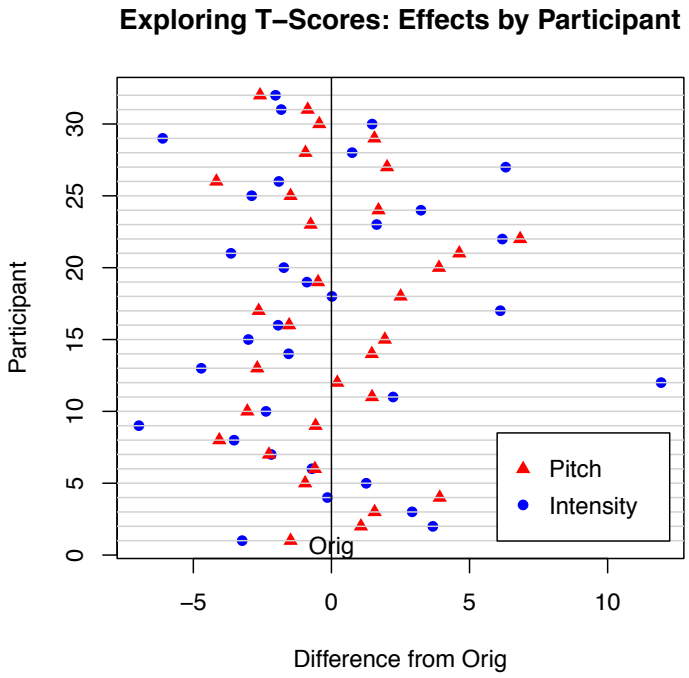

Figure 3.6 Effects plotted by Condition (and by Participant), calculated as T-Scores centered around the Baseline condition. These normalized scores, which present effect directions and sizes on a comparable scale across speakers, illustrate the variation present in speaker performance by treatment.

Indeed, when considering the irregularities associated with acoustic-convergence to speech in previous works, there is little reason to expect entrainment to background noise would be any more consistent, nor that any such effects would have to be linear in nature. Keeping in mind that t-values greater than 2.0 are typically regarded as significant though, even with the conservative values generated through the unpaired t-tests it appears that participants are being reliably influenced by the signal manipulations. We see that values are generally greater than +/-2, however differences by condition – or even within condition, across participants – are not always in the same direction. That is, some participants converge while others diverge, and some do both (across conditions). These transformed data alone, however, do not show any clear trends. And because much of the social information collected is at least somewhat collinear, linear regression alone may not be the most effective way to predict how participants will be affected by manipulations. Use of all predictors could risk over-fitting models. It would make sense, then, to next explore alternative forms of analysis which may shed light on which specific variables are relatively more important when predicting participants' convergent/divergent behaviours.

Having supported that speakers seem regularly influenced by background music it makes sense to next explore which predictors best explain the data, and whether or not how a participant responds to one manipulation can be used to predict performance in another condition. The following analyses involve two methods to aid in answering these questions, first exploring *Random Forests* (e.g., Breiman, 2001; Strobl, Hothorn, and Zeilis, 2009) with the intention of using outputs from those forests to next inform *mixed effects models including polynomial functions* (once again using restricted cubic splines, available through the RMS convenience package in R). Before

getting into the details of each sub-analysis, though, it would be wise to first say a little about random forests in statistics – what they are, and how they work.

Random forests are noted for their ability to deal with 'small n large p' problems, complex interactions, and even highly correlated predictor variables (Strobl, Boulestiex, Kneib, Augustin, and Zeileis, 2008). This method involves a form of machine learning, where data are subset randomly and any specified predictor-variables are tested iteratively for their relative importance in explaining the data/predicting outcomes. Tagliamonte and Baayen (2012) explain that random forests:

> "…work through the data and, by trial and error, establish whether a variable is a useful predictor. The basic algorithm used by the random forests constructs conditional inference trees. A conditional inference tree provides estimates of the likelihood of the value of the response variable… based on a series of binary questions about the values of predictor variables… The algorithm works through all the predictors, splitting (partitioning) the data into subsets where justified, and then recursively considers each of the subsets, until further splitting is not justified. In this way, the algorithm partitions the input space into subsets that are increasingly homogeneous with respect to the levels of the response variable."

Moreover, Strobl et al., (2008) explain how random forest models are so powerful when predicting/simulating data, adding that:

> "In random forests… an ensemble of classification trees is created by means of drawing several bootstrap samples or subsamples from the original training data and fitting a single classification tree to each example. Due to the random variation in the samples and the instability of the single classification trees, the ensemble will consist of a diverse set of trees. For prediction, a vote (or average) over the predictions of the single trees is used and has been shown to highly outperform the single trees: by combining the prediction of a diverse set of trees, [this method] utilizes the fact that classification trees are instable but on average produced the right prediction."

Thus, where dependency trees are optimized locally and recognize value in predictors from the top down, random forests are optimized globally through the generation of many dependency trees using randomized subsets of the data. The relative importance of predictors is therefore defined from the bottom up by optimizing across all trees in the 'forest'. Outputs from random forests are typically summarized as *Variable Importance Plots,* which present all predictor variables tested on a relative scale of importance as distanced from a zero-contribution value. Note that Strobl et al., (2008) maintain "the absolute values of the [variable importance] scores should not be interpreted" – these model outputs are meant only to show relative contributions when explaining/predicting the data, where positive values (above zero) contribute to better understanding the data.

Variable importance plots by Condition are available for the present data as Figure 3.7. Data were fit by condition to generate each plot, where t-values for pitch-data (n = 32) and intensity-data (n = 32) were tested separately. The choice for separate testing was made because it was unclear whether or not there was any correlation between performance across tasks (cf. Figure 3.6). Keeping in mind that any values above "0" are considered viable contributors, we see that HoursMusicPerDay, PresentationOrder (of conditions), and ChooseMusicWhileWork were found to best predict responses in the pitch condition. However, when predicting data from the intensity condition, it appears that PresentationOrder, Musical experience, HoursMusicPerDay, and ChooseMusicWhileWork were found to be most reliable.

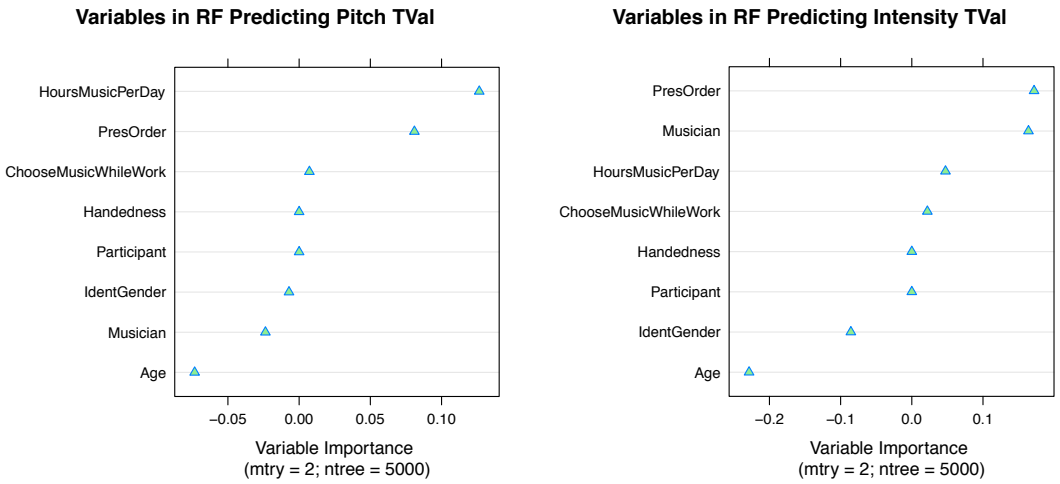

Figure 3.7 Variable Importance Plots by Condition, showing the relative importance of all variables tested. These models both included the parameter settings "ntree = 5000" (where 5000 trees were grown within the forest) and "mtry = 2"(which limits the number of variables tested for contributions at every potential split in the data to 2 of the total number being tested). Strobl et al., (2008) argue that relatively low mtry values result in more diverse trees, and therefore more reliable models.

In order to explore the viability of these predictors across conditions, the random forest models from one condition can be used to generate predicted-data that can then be tested against known data from the other. Specifically, through simulated data we can test whether or not convergence in one condition predicts convergence in another; for example, using the *predict* function in r, we can train models on t-values from one condition (e.g., Pitch) and then use that model to predict t-values for the other (i.e., Intensity). These predicted observations can also be tested for correlations with the actual t-values (from the condition just predicted, in this example the intensity-data) to show how well the model accounts for the data across conditions, thus testing a relationship between performance in the two conditions. These simulations were run using only Pitch t-values OR Intensity t-values to predict those of the other condition. Scatter plots with fitted regression lines and correlation coefficients are available as Figure 3.8 – these figures plot actual t-values on the X-axis against predicted values on the Y-axis. Note that predictions based on both the Pitch and Intensity data result in relatively strong correlations and are highly significant.

It was possible, however, that these high correlations were in fact the product of control data shared across observations (which forms part of the t-values). It would be wise then to test whether distributions would randomly

create effects due to an inherent relationship between pitch and intensity, where effects born through the random forests are actually artifactual of this relationship. One way to establish the reliability of these correlations is to break the link between control and test observations and recalculate the t-values before testing for correlations. To this aim, observations within each treatment were shuffled somewhat (i.e., within gender) and t-values were regenerated. New predictions were made using the semi-randomly sampled data, which resulted in significant decreases in predictability (Figure 3.9 – actual t-values on the X-axis, shuffled t-values on the Y-axis). It should be noted that correlations involving the shuffled/recalculated t-values are still quite strong, indicating that much of the observed correlations are not driven by participants behaving regularly, but instead are driven by artifacts in the data. Though it would have been more straightforward to directly test for correlations between t-values across conditions, incorporating predicted data based on random forest models in this test further takes into account the predictors recognized as important contributors while testing for correlations in participants' performance across conditions.

While not wholly predictive, we do see in this context that a participant's performance in one condition can reveal at least some important information about how they may respond in other treatments – though, correlations do not disappear when data are shuffled, further indicating an inherent link across condition-specific observations. Simply, these tests indicate that there is systematicity within the data that generally connects performance across conditions for participants, in part due to regularity of behaviour and in part due to shared information across conditions through the t-values. These reduced but persistent correlations remaining after data had been shuffled may perhaps be the product of restrictions imposed upon the (semi-)random sampling (i.e., within both gender and condition), though it seems likely that this correlation is at least in part the product of how the t-values were calculated for both treatments. Through these tests, it appears that performance across conditions is only marginally related.

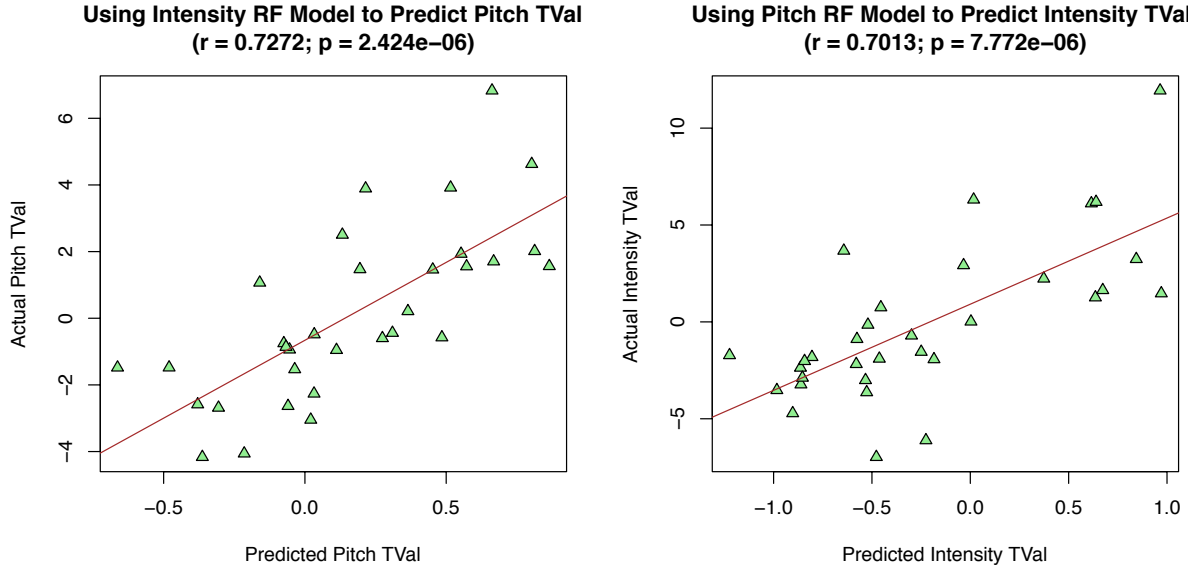

Figure 3.8 Scatterplots and regression lines showing predictions vs. real data based on the random forest models (Left: Predictions based on Intensity values; Right: Predictions based on Pitch Values)

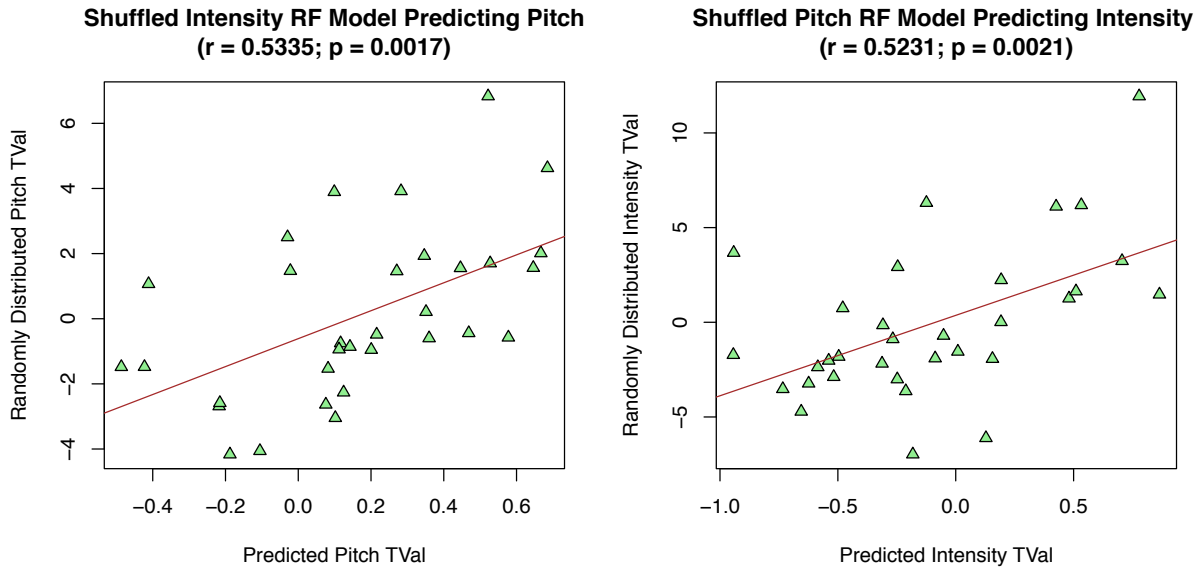

Figure 3.9 Scatterplots and regression lines showing predictions vs. real data based on the random forest models, where data points were mismatched within Condition and within Identified Gender (Left: Predictions based on Intensity values, Right: Predictions based on Pitch values). The relationship between performance in one treatment vs. the other appears to be quite complex, as the mismatch results in decreased – but still reasonable – correlations.

Now that the relative importance of all available predictors has been assessed, those recognized as important contributors can be used to feed mixed effects models (Dilts, 2013; Tagliamonte and Baayen, 2012). In the present case, we have seen (in Figure 3.7) that considering the number of hours spent listening to music per day, the presentation order of the treatments, and whether or not a subject chooses to listen to music while doing cognitively demanding work has resulted in the most robust predictions of voice-pitch convergence/divergence. Similarly, we find that considering the presentation order of treatments, previous musical training, the number of hours spent listening to music per day, and whether or not a subject chooses to listen to music while doing cognitively demanding work provided the most reliable predictions when assessing performance in the Intensity condition. One downfall of random forests, however, is the fact that outputs do not explain exactly *how* these variables may be influencing outcomes. Specifically, interactions are taken into account somewhat through the forests, though little can be known about the specifics of such interactions through model outputs. It would therefore be sensible to feed the appropriate variables into mixed effects models (by Condition) to learn more about how these variables may influence outcomes.

More descriptive exploration would be wise before diving into the mixed effects modeling though, where checking density plots for certain variables recognized as important in the forests but are likely to be correlated with each other can ensure (1) that various predictors are actually serving different purposes (that is, providing unique information) within the modeling process, which (2) avoids over fitment of the models by avoiding redundant predictors. Therefore, this descriptive sub-analysis will also ensure the data (3) are meeting assumptions of the mixed effects modeling with regard to correlated predictors, and (4) that all factorial levels in the data are well represented

as *actual* observations (as opposed to predicted data or gaps), where gaps in the data can result in less reliable modeling and falsely inflated effects. Two such comparisons can be found below (Figure 3.10) in HoursMusicPerDay + Musicianship, and in HoursMusicPerDay + ChooseMusicWhileWork:

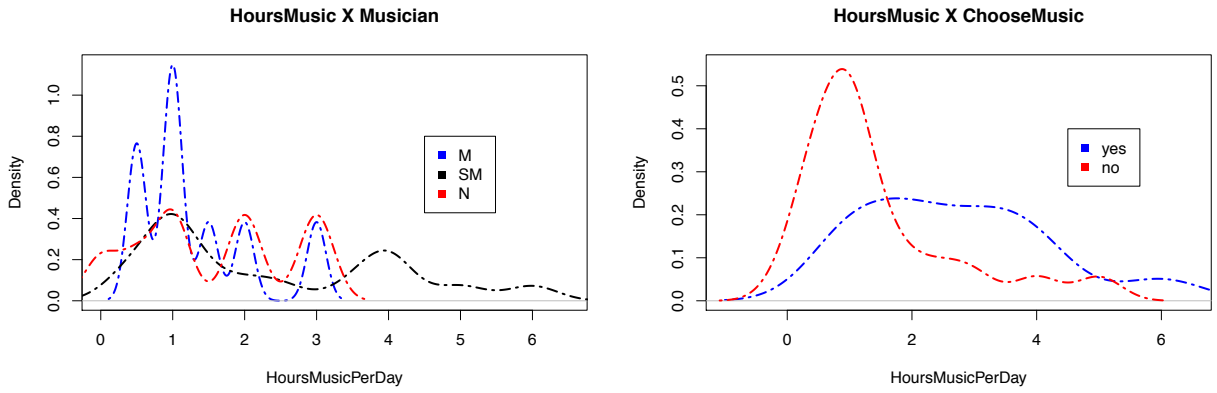

Figure 3.10 Density plots comparing the distribution of data across categories. (Left: Distributions of HoursMusicPerDay by level of Musicianship, Right: Distributions of HoursMusicPerDay by whether or not a participant chooses to listen to music while doing cognitively demanding work).

The above density plots express useful information about trends in the data, as well as the reliability of those trends. For example, in Figure 3.10 (Right) we can see very clearly that speakers who choose not to listen to music while working through cognitively demanding tasks also generally listen to very little music per day. Conversely, while there is some overlap, it appears that speakers who choose to listen to music while working through such tasks show much more even distribution over the space expressing 'hours music per day', though also self-report as listening to relatively more music on average. These two variables appear somewhat related, however it seems they mostly offer distinct information when used to model the data. When exploring the plot on the left of Figure 3.10 however, we see less distinct patterning when considering HoursMusicPerDay in light of the speaker's previous musical training. Two trends immediately draw attention within the plot: First, and counter to intuitions, it seems *musicians generally listen to less music than both the non-musicians and the speakers with some musical training*. Secondly, we also see that *the listeners with some musical training appear to listen to more music than the other two groups*. However, while some of the musicians report they listen to very little music, many within this group are patterning exactly like the non-musicians. In fact, while all three groups are well represented at the low end of the spectrum, these two variables may be difficult to tease apart due to relatively similar patterning across the space of HoursMusicPerDay. Exploring these plots, therefore, informs how models might be shaped and the summaries of effects interpreted. I do believe all of these variables may be useful in the modeling that follows, though these relationships across variables should be considered when interpreting model summaries. Specifically, it seems a speaker's degree of musicianship is somewhat related to/confounded with HoursMusicPerDay in these data, and any effects related to either of these variables predicted through the mixed effects modeling might be considered with this relationship in mind.

In brief summary, it appears the above analyses support these differences in participant responses (as observed through the t-value transformation) as largely systematic, and therefore not random in nature. We have seen that observations from one condition can be used to predict responses from another to some extent, though part of this correlation across conditions appears the product of a shared baseline condition being used when generating t-values for both treatments. We have also identified a select group of variables which are believed to maintain relative importance as viable predictors when explaining variation in the data – these predictors are somewhat similar, though not identical across conditions. Having supported responses as non-random in nature, we might next investigate the social and experiential aspects that may contribute to shaping any observed effects.

3.10.2 If Performance is Not Random, then Which Social Variables Best Explain Responses?

In the following models Mean Pitch by item and Mean Intensity by item served as dependent variables (in separate analyses). Backward-stepwise model selection was used to define fixed effects structures, where predictor variables were removed one by one and tested for significant contributions to the model fit. All predictors that significantly improved the model fit were retained (with a few exceptions described below). All two-way interactions involving variables recognized as important through random forests (i.e., through any VIP score above "0") were also tested, with the addition of IdentGender as a control variable given expected differences in voice-pitch and motivations in previous literature. Models were compared progressively using ANOVA, where models with lower AIC scores were preferred. Random intercepts were included for Participant and Item, as were random slopes by Condition; the inclusion of random slopes in this context allows for the slope of effects to differ by Condition, resulting in more precise identification of trends in the data. The optimizer was set to "bobyqa". Because the outputs from the random forests differed so dramatically from the effects structures in all previously discussed linear modeling in EXP.1, it was presumed that these effects may in fact not be linear in nature. All following models were therefore initially fit with a restricted cubic spline using the rcs function (available through the RMS package (Harrell, 2018)) which allows for some degree of alinearity in predicted effects.

Across all conditions, the dataset included 5,184 observations.

3.10.2.1 Post-Hoc Re-Analysis: PITCH

In order to parallel analyses in EXP.1 as closely as possible while modeling Mean Pitch (as calculated by item), data were subset by Condition where only observations collected during the Pitch and Original conditions were included in this analysis (n = 3,456 observations). All possible two-way interactions between HoursMusicPerDay, Condition, PresOrder (describing the presentation order of conditions in a session, somewhat analogous to Block in EXP.1 – this variable is meant to facilitate testing for persistent effects and presentation-order effects which were both confounded in the first experiment), ChooseMusicWhileWork, and IdentGender were tested and retained if significant.

Some points of note before describing results: During the modeling process fitting to the restricted cubic spline resulted in relatively high VIF (Variance Inflation Factor) scores when testing the severity of any potential multicollinearity. Therefore, this function was removed and the model once again fit to a simple linear function. Additionally, the HoursMusic variable was recognized as a significant predictor during modeling, though its inclusion also resulted in high VIF scores. After plotting the interaction of primary importance (HoursMusic X Condition) a split in performance was recognized at 1 hour or less vs. more than one hour; this variable was therefore recast as binary and split accordingly. The one-hour mark (inclusive) also divides the participant distribution exactly in half, so both sides of this split are equally represented. Making these changes brought all VIF scores into an acceptable range. Finally, while PresOrder was found to contribute significantly to the model through an interaction with Condition, these effects were not easily interpreted and were also largely based on predicted data. That is, as a result of the overly-complex model the data were spread too thin and this effect appeared unreliable. PresOrder was therefore removed from the final model, which is available below as Table 3.1.

	Estimate	Std.Error	t-value
(Intercept)	182.1132	4.9936	36.469
BinaryHoursMoreThanOne	-1.0365	7.765	-0.133
ConditionPitch	**2.8735**	**0.6976**	**4.119**
ChooseMusicWhileWorkY	**-26.566**	**10.8254**	**-2.454**
IdentGenderMALE	**-72.8689**	**6.3081**	**-11.552**
BinaryHoursMoreThanOne:ConditionPitch	**-4.8843**	**0.9867**	**-4.95**
BinaryHoursMoreThanOne:ChooseMusicWhileWorkY	**42.7576**	**13.6122**	**3.141**

Table 3.1 The final model predicting mean F0 in the Pitch condition, using predictors from the random forest modeling. Significant effects have been bolded for convenience (as t = +/- 2.0 or greater).

As can be seen above in Table 3.1, main effects were observed for Condition (Est. 2.8735, t = 4.119), ChooseMusicWhileWork (Est. -26.566, t = -2.454), and for Identified Gender (Est. -72.8689, t = -11.552). We also find a significant interaction between the binary HoursMusicPerDay variable (BinaryHours) and Condition, where productions from speakers listening to more than one hour of music per day exhibit significantly lower voice-pitch in the lowered-pitch condition than in the Baseline treatment i.e., convergent behaviour (Est. -4.8843, t = -4.95). Moreover, it appears that productions from speakers who listen to one hour or less of music per day exhibit significantly higher F0 in the lowered-pitch treatment – or divergent behaviour – when compared to productions from the baseline. This interaction is plotted below as Figure 3.11, with the binary measure of hours music per day on the X-axis, Pitch in Hz on the Y-axis, and different coloured lines representing different Conditions.

At this point, though, it makes sense to subset the data by groups comprising the BinaryHours condition to ensure differences across conditions are in fact both significantly different. Sub-setting in this way left 1728 observations for the MoreThanOne hour group and 1728 observations for the OneOrLess group. Removing all terms involving the BinaryHours variable (though, otherwise maintaining the same fixed effect, random effect, and random

slope structures) indicates significant main effects of Condition for both groups (OOL: Est. 2.81, t = 3.695; MTO: Est. -2.0612, t = -3.279), supporting convergence for speakers who listen to MoreThanOne hour of music per day and divergence for those who listen to OneOrLess hours of music per day.

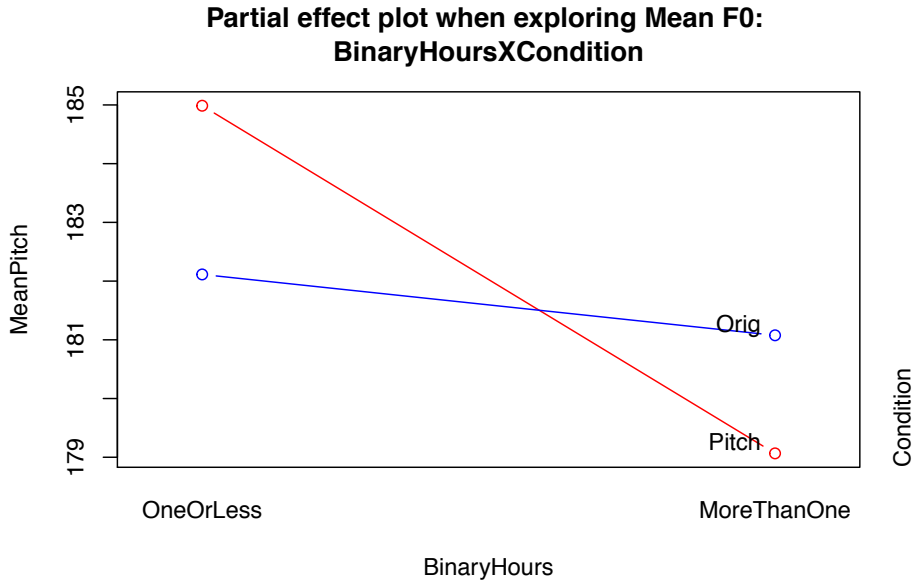

Figure 3.11 Partial effect plot showing the interaction between HoursMusicPerDay and Condition, where participants listening to 1 hour or less of music per day show signs of divergent behaviour, whereas those who listen to more than one hour per day appear to converge with the lowered-Pitch treatment.

We also find BinaryHours interacting with ChooseMusic. Here, we see that productions from speakers who listen to more than one hour of music per day and also choose to listen to music while doing cognitively demanding work tend to have significantly higher voice-pitch than those who do not Choose Music while working (Est. 42.7576, t = 3.141). Conversely, we find that productions from speakers who listen to one hour of music per day or less and do not choose to listen to music while working exhibit significantly lower fundamental frequency than those who do choose music while working.

Though, it makes sense to subset the data once again by groups within the BinaryHours condition to ensure the by-condition differences are in fact both significant. Subsetting in this way left 1728 observations for each of the MoreThanOne hour and the OneOrLess groups. Removing all terms involving the BinaryHours variable, while otherwise maintaining the same fixed effect, random effect, and random slope structures indicates that the relatively convergent behaviours exhibited by the OneOrLess group who ChooseMusic while working remains significant i.e., a main effect of ChooseMusic (Est. -25.6647, t = -2.654). Conversely, the relatively divergent behaviour seen in speakers who listen to MoreThanOne hour of music per day and also ChooseMusic does not reach significance (Est. 16.1565, t = 1.698). It seems likely that much of this effect would be the product of inter-subject differences, which could be measured equally well by the random intercept. A partial effect plot for the BinaryHours X ChooseMusic interaction is available below as Figure 3.12 with the binary measure of hours music per day on the X-axis, Pitch in

Hz on the Y-axis, and different line colours indicating whether or not the speakers choose to listen to music during cognitively demanding work.

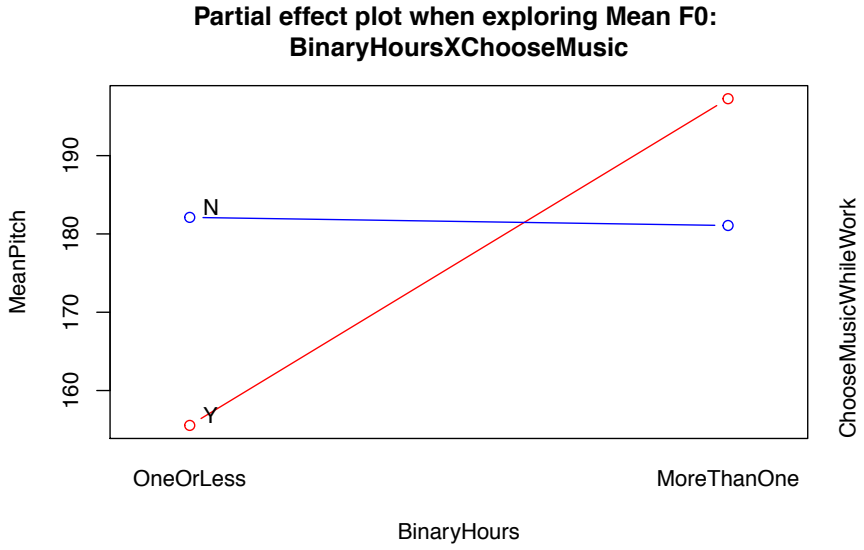

Figure 3.12 A partial effect plot showing the interaction between HoursMusicPerDay and ChooseMusicWhileWorking. In this plot, we see that productions from speakers who listen to more than one hour of music per day and also choose to listen to music while working have higher mean fundamental frequency than those from speakers who do not choose to listen to music (though this difference does not reach significance). Also, we see that productions from speakers who listen to one hour of music per day or less and choose to listen to music while doing cognitively demanding work exhibit significantly lower F0 than those who do not choose music while working.

3.10.2.3 Post-Hoc Re-Analysis: INTENSITY

This analysis was also conducted using mixed effects modeling, though predicting Mean Intensity as calculated by item. Random effects were included for Participant and Item, as were random slopes by Condition. The optimizer was set to "bobyqa". Again, to parallel analyses in EXP.1 data were restricted by Condition to only observations collected during the Intensity and Original conditions (n = 3,456 data points). All possible two-way interactions between HoursMusicPerDay, Condition, PresOrder, ChooseMusicWhileWork, Musical Experience, and IdentGender have been tested and were generally retained if significant (exceptions described below).

Much like in the Pitch analysis, fitting to the restricted cubic spline resulted in high VIF scores. As a result, these models were also fit to a simple linear function. The HoursMusic variable was again troublesome with regard to high VIF scores in this analysis, though use of the binary HoursMusic variable (i.e., 1 hour or less vs. more than one hour) brought VIF scores into an acceptable range. Moreover, while inclusion of the PresOrder variable resulted in a significant interaction between PresOrder and Condition, this complex analysis spread the data very thin and

much of this effect was again based on predicted data (i.e., empty cells). PresOrder was removed from the final model as a result, and the model summary is available below as Table 3.2.

	Estimate	Std.Error	t-value
(Intercept)	73.14347	1.51822	48.177
BinaryHoursMoreThanOne	-1.11277	1.26899	-0.877
ConditionInt	-0.10097	0.12868	-0.785
MusicianSM	-1.73303	1.56031	-1.111
MusicianN	-2.74728	1.84941	-1.485
IdentGenderMALE	-0.51535	1.37805	-0.374
BinaryHoursMoreThanOne:ConditionInt	**-0.69132**	**0.10738**	**-6.438**
ConditionInt:IdentGenderMALE	**0.24082**	**0.11715**	**2.056**
ConditionInt:MusicianSM	**0.78609**	**0.13234**	**5.94**
ConditionInt:MusicianN	-0.03488	0.15652	-0.223

Table 3.2 The final model predicting mean Intensity, using predictors from the random forest modeling. Significant effects have been bolded for convenience (as t = 2.0 or greater).

The model indicates a significant interaction between the binary measure of HoursMusicPerDay (BinaryHours) and Condition when predicting mean intensity. Once again, it appears that productions from speakers who self-report as listening to more than one hour of music per day tend to speak with significantly lower intensity during the lowered-Intensity treatment than in the Baseline condition (Est. -0.69132, t = -6.438), and those who listen to one hour or less per day exhibit relatively higher intensity. Recognizing that this partial effect describes musicians specifically (set as reference from a three-level factor), it makes sense to plot this interaction with different levels of musicianship set as reference to ensure behaviour is similar across groups. This approach indicates that both musicians and non-musicians who listen to more than one hour of music per day exhibit convergent trends during the lowered intensity treatment; musicians and non-musicians who listen to one hour or less per day however, both show no effect of the manipulation. A plot of the musicians is available below in Figure 3.13 (Left), and is meant to represent both the musician and non-musicians (these plots appear nearly identical). A plot for the speakers with some musical training is also available in Figure 3.13 (Right), who show the opposite trend – SMs who listen to one hour or less are exhibiting significantly divergent behaviour in the lowered-intensity treatment, whereas those who listen to more than one hour of music per day seem unaffected by the manipulation. These plots together suggest there may be an important three-way interaction between BinaryHours, Condition, and Musician.

However, it makes sense to next test the reliability of these partial effects for each group by subsetting the data as in the pitch-based analyses. Restricting the data by levels of musicianship and levels of BinaryHours allows for testing of significant effects of convergence/divergence across conditions; subsetting in this way leaves 864 observations for musicians (OneOrLess: 540, MoreThanOne: 324); 1728 observation for the SMs (OneOrLess: 756, MoreThanOne: 972); and 864 observations for the non-musicians (OneOrLess: 432, MoreThanOne: 432). Fitting

models to these restricted datasets explores the possibility of main effects of Condition within each group, thereby testing the effectiveness of the treatment for each sub-group. In these models all terms related to BinaryHours and Musician were removed, while otherwise maintaining the fixed effect, random effect, and random slope structures. These models indicate that musicians who listen to more than one hour of music per day are on the cusp of convergence (Est. -0.376, t = -1.938), whereas musicians who listen to one hour or less show no effect of the manipulation (Est. -0.1482, t = -1.087). Non-musicians who listen to more than one hour of music per day exhibit convergent behaviour (Est. -0.3637, t = -2.421); though, interestingly, non-musicians who listen to one hour or less also appear to converge with the lowered intensity manipulation (Est. -0.5381, t = -4.412). Conversely, SMs who listen to one hour of music or less per day exhibit significantly divergent behaviour in the lowered-intensity treatment (Est. 1.2307, t = 9.487), while SMs who listen to more than one hour of music per day appear to converge with the lowered intensity treatment (Est. -0.23968, t = -2.636). With regard to the behaviours observed in the Pitch analyses, it appears the SMs are showing similar effects, whereas the non-musicians appear more prone to convergence regardless of listening habits. The musicians, on the other hand, are the least likely to show any effect of the intensity-based manipulation (cf. EXP.1), though are showing convergent trends where expected.

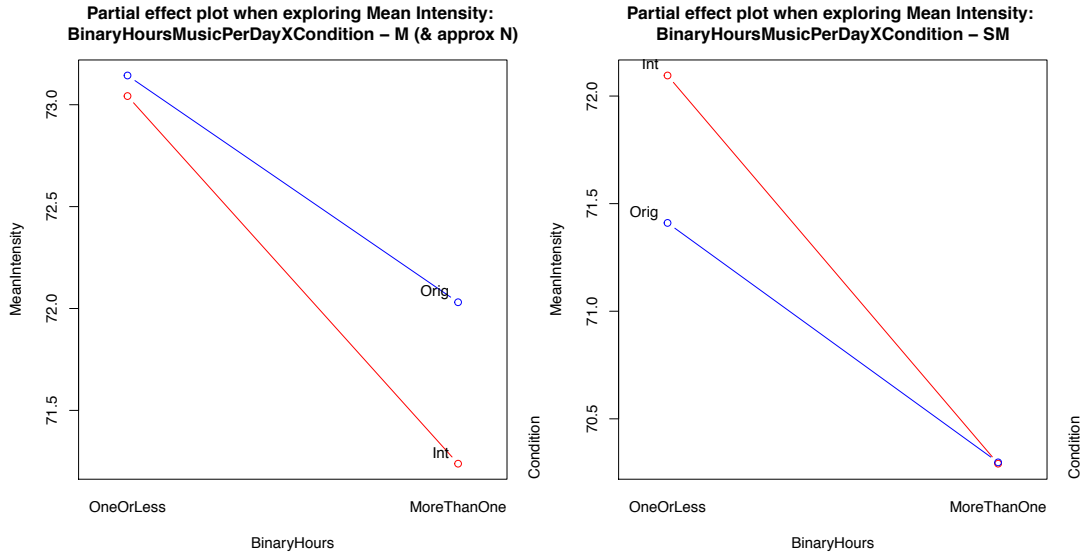

Figure 3.13 Partial effect plots showing the interaction between HoursMusicPerDay and Condition when predicting mean intensity. On the left we see that both musicians and non-musicians listening to 1 hour or less of music per day show signs of divergent behaviour, whereas those who listen to more than one hour show no effect of the lowered-Intensity treatment; however, on the right we see SMs showing the opposite trend. It appears speakers with some musical training who listen to one hour or less of music per day exhibit divergent behaviour, while SMs who listen to more music appear to show no effect of the manipulation.

We also find a significant interaction between Condition and Musician, where productions from the speakers with Some Musical training appear to be significantly louder than the musicians in the lowered-intensity condition (Est. 0.78609, t = 5.940); productions from the Non-musicians were not significantly different from the musicians (Est. -0.03488, t = -0.223). Releveling with the SMs as the reference in order to test for any significant

difference between the Non-musicians and the SMs shows that these productions are also significantly different (Est. -0.8210, t = -6.151). It appears that in this interaction musical training is useful in predicting the magnitude of divergence, though all talkers would be predicted to diverge through this partial effect (as seen in Figure 3.14 with Musical experience on the X-axis, mean intensity on the Y-axis, and different coloured lines indicating which condition is being predicted).

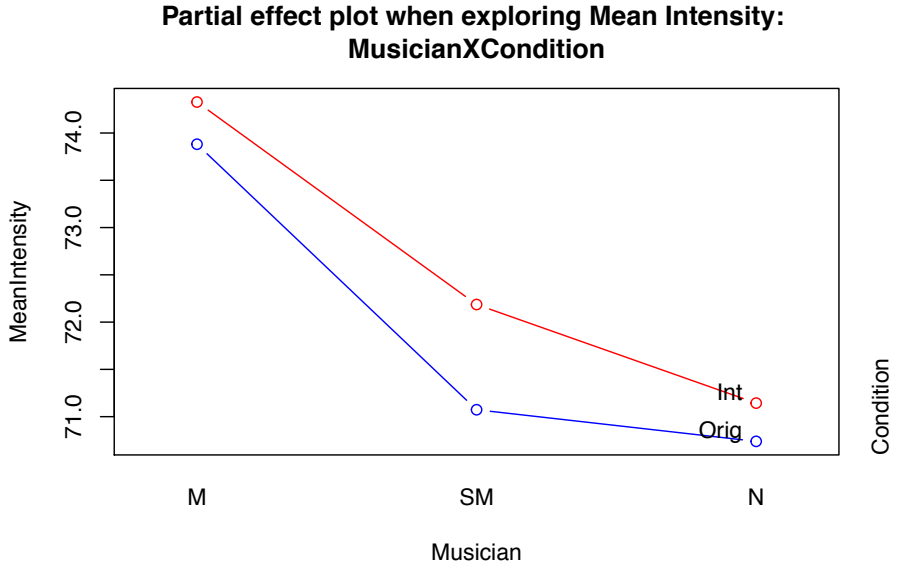

Figure 3.14 A partial effect plot showing the interaction between Musical experience and Condition when predicting mean intensity. We see that all forms of musical experience indicate speaker divergence by way of intensity, however the speakers with some musical training appear to show a relatively larger divergence than the other groups.

However, once again it would be wise to subset the data by musicianship in order to test the reliability of these effects within each group. Subsetting the data in this way leaves 864 observations for musicians, 1728 observation for the SMs, and 864 observations for the non-musicians. Removing all terms involving Musicianship and all interactions, but otherwise retaining identical model structures indicates that all groups retain significant effects of Condition (M: Est. -0.2439, t = -2.184; SM: Est. 0.40279, t = 5.173; N: Est. -0.44689, t = -4.634), where effects and directions of those effects fit well with findings from the above BinaryHours X Condition analysis. Once again it appears both the musicians and non-musician exhibit convergent behaviours in the lowered-intensity condition, whereas the SMs appear to be diverging.

Finally, we find a significant interaction between Condition and IdentGender where productions from speakers who identify as male tend to be significantly louder in the lowered intensity condition than the participants who identify as female (Est. 0.55077, t = 4.223). While both groups are predicted to produce speech of relatively higher intensity in the lowered-intensity condition (that is, when compared to the unaltered baseline), male speakers are predicted to show a relatively larger discrepancy between the two conditions – however, this larger gap appears to be the product of male speakers' productions being relatively more quiet in the unaltered condition as opposed to

relatively louder when responding to the intensity-based treatment. This interaction is available below as a partial effect plot in Figure 3.15 with Identified Gender on the X-axis, mean intensity on the Y-axis (in dB), and different coloured lines expressing the condition being predicted.

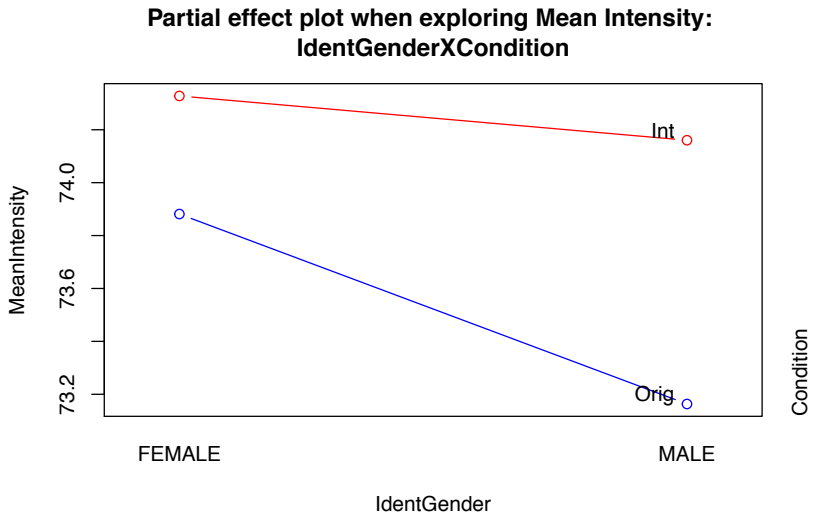

Figure 3.15 A partial effect plot showing the interaction between Identified Gender and Condition when predicting mean intensity. We see that both groups of speakers are predicted to speak relatively more loudly in the lowered-intensity condition, though male speakers are predicted to speak relatively more quietly in the unaltered condition, resulting in a relatively more sizable gap between conditions.

These effects can be corroborated through another subsetting of the data, though this time by way of the IdentGender variable. Restricting the data in this way leaves 2268 observations for the female speakers and 1188 for the male speakers. All terms involving the IdentGender variable were removed, as were all interactions and terms involving musical training for ease of interpretation – however, a random effect for Musician was added to the previous random effect structure (Participant and Item) to account for a known influence of musicianship. Random slopes by Condition were also still included. Modeling in this way indicates a significant effect of Condition for speakers identifying as male, who were predicted to be significantly louder in the lowered intensity condition than in the baseline condition (Est. 0.25379, t = 2.548) i.e., divergence. Speakers identifying as female were predicted to be relatively quieter in the lowered-intensity condition, though this difference was not significant (Est. -0.09187, t = -1.436).

3.11 Discussion

While results in the present study did not exactly replicate those of EXP.1, statistical methodologies were refined such that multiple tests in EXP.2 are now providing relatively complementary outcomes. Indeed, some similar pitch- and intensity-based effects appear to have patterned in the data across the two studies; though, these convergent and

divergent effects have primarily been attributed to different predictors across experiments. Where the low participant numbers and time-series data collected during EXP.1 resulted in many analytical issues and unstable analyses, the individual items produced as word lists for EXP.2 and increased participant numbers have allowed for more robust analyses and the avoidance of issues associated with autocorrelation in time series data. For convenience, a summary table of the effects most clearly related to convergent and divergent behaviour observed in EXP.2 is available below as Table 3.3.

Acoustic Dimension	**HoursMusicPerDay**	
	1hour or less	*> 1 hour*
Pitch (was decreased)	Y-inv	Y
Intensity (was decreased)	Musicians: N SMs: Y-inv Non-musicians: Y	Musicians: Y SMs: Y Non-musicians: Y

Table 3.3 A partial summary of results for Experiment 2: "Y" indicates the presence of a significant HoursMusic*Condition interaction, whereas "N" would indicate that no interaction was found. The abbreviation "inv." is again used to identify situations where significant deviation was observed, though in a direction opposite to that predicted given the design of the test condition i.e., an inverted effect. It appears use of the BinaryHours variable results in comparable effects observed across conditions. Recall a significant effect of Musician was observed in the intensity analysis (cf. EXP.1)

Two primary questions are explored throughout the present work: (1) *Is the acoustic variation observed in productions across speakers (and conditions) random?* And, (2) *If performance across conditions does not appear random, then which social variables might best explain how participants respond to the acoustic manipulations?* The immediately preceding analyses brought to light various effects describing how speakers appear to be influenced by ambient music encountered while reading aloud. The discussion that follows aims to interpret those results in a way that addresses these questions.

When considering the first question, regarding whether or not variation observed across productions is random or systematic, through multiple tests and experiments it appears the variation observed is relatively structured and regular – though this question can actually be thought of as two questions. We can ask whether the variation encountered in productions is systematic *within* conditions (that is, are speakers influenced reliably within a condition?), and we can also ask if variation might be systematic *across* conditions (simply, is performance in one condition in any way indicative of performance in another?). I will first discuss the within-condition prospect before moving on to the across-condition possibility.

In Figure 3.6 we see the t-values for each speaker by condition. While these values were generated partially as a data transformation, it is the case that the majority of these values/differences would be considered statistically significant at the ≥ 2.0 criterion. Moreover, such significance is the case despite this transformation involving the more conservative unpaired form of the t-test. Given the relatively high number of observations per condition for

each speaker (n = 54), in addition to use of the conservative test, it seems that these values generated by comparing test-condition observations to those from the baseline are reliably different, and indicate systematicity in performance for speakers within each condition.

When considering potential variability across conditions, the use of random forests and data simulations tell a somewhat similar story. These random forests identified variables of relative importance in separate pitch- and intensity-based analyses. While some variables identified were shared across treatments, the lists were not identical; both recognized the HoursMusicPerDay (BinaryHours) as important, which has proved to be the most telling variable in the process thus far. In other words, BinaryHours was highly significant when predicting performance in both treatments. This connection supports the possibility that performance may be related across conditions.

However, further support is offered through use of the random forest models fit to data from one condition to predict data from the other, and then comparing those predicted data to known observations. This process also shows some connection between participants' performance across conditions. We have seen that simulated data were highly correlated to the known data, and that these correlations were statistically significant. In order to test whether or not these correlations were real or artifacts of a shared control condition involved in generation of the t-values, data were somewhat mismatched and the t-values & correlations recalculated (though, in retrospect, it may have been an error to restrict the sampling process to anything other than Condition during this process – which may have partially maintained the persistent correlation). These recalculated correlations were markedly lower than those previously seen, though still of moderate strength. These results suggest that participants are somewhat reliable in their performance across conditions, though the correlations appear at least partially due to the inherent link between conditions through the t-values (i.e., the shared Control condition involved in generating the t-values for both conditions). Nevertheless, at this point the data support variation observed in productions within Speaker, within Condition, and across Condition as not random in nature, and at least somewhat predictable though variables which explain the data (some of which are shared across conditions), testing predicted data, and the initial t-value transformation.

Next, when answering the second major question regarding which variables may help predict variation in speech production (if variance was found not to be random), the predictors gleaned through random forests resulted in efficient and effective modeling. In fact, nearly all of these predictors were recognized as significant contributors during the mixed effects modeling that followed. However, one issue noted through this use of random forests was the fact that IdentGender was not recognized as an important contributor in the variable importance plots, but was found to be highly significant in both of the subsequent mixed effects models. This oversight does not negate the fact that predictors were identified through random forests and were useful when aiming to explain some of the variance observed in the data, though it does reinforce the analytical strength realized through use of multiple, complementary tests. Some of the variables recognized as important and their related effects will be discussed in detail below, starting broadly with effects related to Condition (and signal manipulations) discussed in light of results from EXP.1.

As suggested earlier in this work, the notion of entrainment to intensity-based variation in background signals appears in some ways much like effects first described by Étienne Lombard (1911). Importantly, literature on the Lombard effect describes different thresholds that have been recognized as best estimates for eliciting Lombard

speech when encountering speech-like background noise vs. non-speech noise. With regard to intensity-based variation encountered in background noise only (recall that there are other acoustic changes which are also associated with the Lombard effect), EXP.1 presents some evidence indicating that participants do in fact converge to intensity levels of background music at and above presentation levels of ~45 dB(A), the known floor for eliciting Lombard speech through non-speech noise (Lazarus, 1986). One may speculate based on this result that background musical noise might be processed cognitively as more like non-speech signals than speech-based signals, where the floor for the latter was recognized by Lazarus (1986) as roughly 55 dB(A) when eliciting this effect. However, the present work does not supply any conclusive information regarding the cognitive mechanisms involved in these processes – only hints for future work. Conversely, results from EXP.2 can more conclusively shed light on how similar or different entrainment effects may be from the Lombard effect. That is to say, if entrainment to intensity is actually just the Lombard effect, then we would predict no convergence to signal-intensity would take place in EXP.2 because the lowered-intensity condition falls 6 dB below the known floor for eliciting Lombard speech.

Put simply, stimuli were again presented at ~45 dB during the baseline condition in the present work, though the intensity-related manipulation in EXP.2 *decreased* intensity levels to ~39 dB(A). Thus, the present manipulation was designed as a means to potentially distinguish the Lombard response from acoustic convergence to ambient intensity – of course, keeping in mind that this work only tests one of the many aspects of speech known to vary in Lombard responses. One point of note with regard to Lazarus' (1986) thresholds identified for eliciting Lombard speech: These are the best estimates currently known. Of course one must recognize that all speakers will differ from one another, even if marginally at times, and that context is likely to further influence how each speaker responds to various conditions. However, the thresholds described by Lazarus were regular enough to disregard the possibility of chance performance, which has been taken into account during the present work. In fact, the article from Lazarus is largely the synthesis and interpretation of multiple other works dealing with various forms of noise at a variety of presentation levels, where the argued thresholds control for certain differences in performance that may be driven by noise-type (to a reasonable extent). That is to say, while I recognize that some speakers may not adhere strictly to these thresholds, it is assumed that the majority of speakers will respond similarly to the trends described by Lazarus.

The fact that statistically reliable intensity-effects of both convergence and divergence were observed when background music was presented below the known Lombard-threshold serves as important evidence showing that either (1) Lombard responses and Entrainment effects are in fact distinct events and processes, or (2) the known floor for eliciting Lombard speech is, in fact, lower than previously believed. It seems likely that certain peripheral acoustic changes noted in Lombard speech research (e.g., Cooke & Lu, 2010) either would or would not be detected in the lowered intensity productions through further acoustic analyses, which could more conclusively relate or distinguish these events/effects – though, such analyses fall outside the scope of the present work. Regardless, the fact remains that like-responses were generally observed for both the pitch- and intensity-based conditions for groups in the current work, where talkers who self-report as listening to more than one hour of music per day converged to both the pitch and intensity of background music. Moreover, it also seems that participants who estimate they listen to one hour of music per day or less reliably inverted this effect cf. figs. 3.11, 3.13 (though, recall the influence of

musicianship on responses during the intensity condition). Keeping in mind that most Musicians in EXP.2 report listening to very little music (Figure 3.10, left), as well as the fact that it was the Musicians in EXP.1 who were found to diverge in the pitch treatment (see Table 1.3), these results across studies appear to be somewhat more coherent than initially thought. Divergence in the SMs could make sense if the majority of these speakers had relatively more musical training, though it is unclear at present why the musicians showed no effect in this context.

Importantly, the influence of listening habits (as measured through the BinaryHours variable) alone instead of musicianship does not completely explain participant behaviours. While we see similar trends in responses across studies when comparing EXP.1 and EXP.2 through inclusion of BinaryHours, musicianship appears still to be playing an important/significant role, at least in the Intensity condition. The observed effects seem not simply explained by *either* musicianship (EXP.1) *or* BinaryHours (EXP.2), though considering both of these variables together can help better understand participant responses.

Looking at some of the data distributions from both studies can draw attention to commonalities between the degrees of musicianship and speaker listening habits. For example, perhaps the one-hour-or-less split point that suits data from EXP.2 might better be treated as an estimate than a hard line, which is realized within some range based on the summed experience of the group (forms of experience gained through both musical training and listening habits). If this were the case, we would predict that the crossover point observed in EXP.1 would also be in the general vicinity of 1 hour, though not necessarily the exact same spot. Plotted below (as Figure 3.16) are histograms showing distributions of HoursMusicPerDay from EXPs 1+2; note that the distribution is cut in half at one hour per day for EXP.2, and is split at two hours per day for EXP.1 (where HoursMusic is listed on the X-axis, and frequency on the Y-axis).

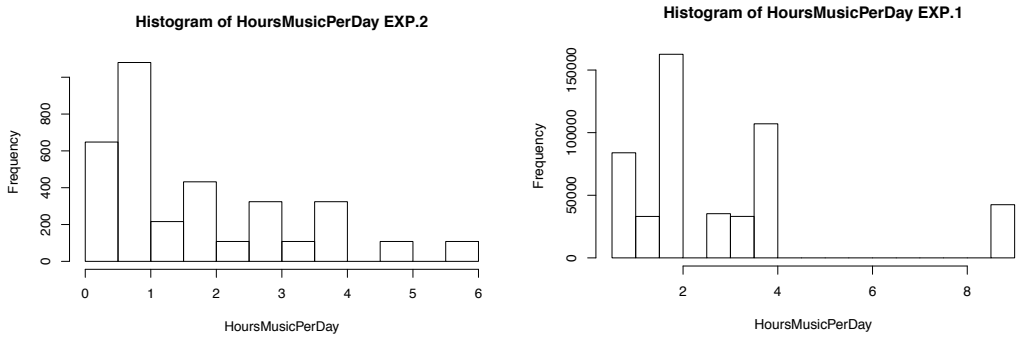

Figure 3.16 Histograms illustrating the distributions of HoursMusicPerDay representation in EXP2 (left) and EXP1 (right). Where EXP2 representation is split at one-hour or less, representation in EXP1 is split at 2-hours or less. Provided distributions of HoursMusicXMusician are comparably similar, these plots lend support to the theory that perhaps some currently unknown drive related to both Musician and HoursMusic may best predict convergence/divergence to ambient music.

Next, if distributions of HoursMusic by Musician (i.e., degree of musicianship) are comparably similar, then together these plots would support both musical experience and musical listening habits as contributing to the observed effects to some degree, and perhaps interacting at times to shape the realizations of these effects. Such

similarities could at least partially explain why the two variables have so far most effectively predicted convergent and divergent behaviours in speech production. Density plots of these distributions are available below as Figure 3.17, again with HoursMusic listed on the X-axis and frequency on the Y-axis. I draw attention to the fact that in both plots we find the Musicians are best represented at or below the cut-off (EXP2: ≤ 1 hours; EXP1: ≤ 2 hours) and both the Non-musicians and SMs are best represented at or above their respective cut-offs. It therefore seems that the differences in results when comparing EXP.1 to EXP.2 *may* be less problematic than initially thought because the different variables predicting these similar effects (across studies) appear to share some important relationship in this context. With this as the case, it seems likely that perhaps neither Musicianship nor HoursMusic alone are likely to be *the best predictor* for convergent and divergent behaviours in this context. Indeed, they both appear to have some predictive power and seem to be involved together in some dynamic relationship, or are perhaps related to some unknown variable which may in fact be the primary motivator of these effects.

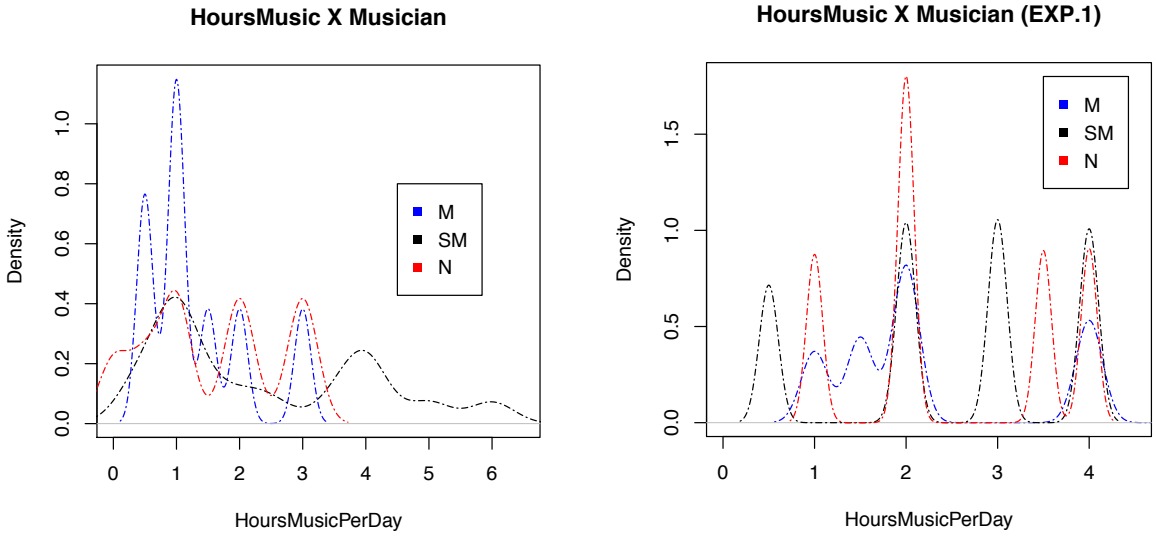

Figure 3.17 Density plots of HoursMusicPerDay given previous musical training in EXP.2 (left) and EXP.1 (right). EXP.2 has been re-plotted here for ease of side-by-side comparison. While these plots are not identical, there are some notable similarities in patterning. We see that Musicians are best represented at or below the HoursMusic distribution cut-offs, whereas the Non-musicians and participants with Some Musical training are best represented at or above those cut-offs.

Attention must be drawn though to certain similarities and differences observed in the EXP.2 intensity-analysis driven by musicianship (i.e., differences when compared to the EXP.2 Pitch analysis, and when compared to EXP.1 Intensity analysis). In the way of similarities we see that the SMs are adhering to the general trends observed in the Pitch analysis in EXP.2, where BinaryHours seems largely a viable predictor for similar behaviour across treatments. Moreover, we also see considerable similarity in performance in the intensity-based treatments when comparing performance across the two studies. Recall that all speakers were found to converge with intensity manipulation in EXP.1; this was nearly the case in EXP.2 as well. Keeping in mind the split between musicianship and BinaryHours groups, all grades of musicianship in the MoreThanOne hour group exhibited convergent behaviour

to intensity variation, as did the non-musicians who listened to OneOrLess hours of music per day. Differences arise when we consider the musicians who listen to relatively little music showing no effect, and the SMs who were found to diverge in this condition.

These differences could perhaps be the product of the altered methodology (i.e., continuous vs. global manipulation), a low-powered analysis, or the low participant numbers in EXP.1 – but I do not believe this to be the case, or at least not wholly. It seems possible that differences in *some* of the SM responses may perhaps have been missed in EXP.1 while conflating the SMs with the non-musicians. If this were the case then we would expect that the magnitude of the intensity-based convergence effects experienced by non-musicians in EXP.1 would have been reduced somewhat, due to the divergent SMs who report listening to OneOrLess hours of music per day. It also seems very possible that some of these differential effects were missed in EXP.1 due to the fact that HoursMusic was not included in the analysis of that study. Recall that final models in EXP.1 were stripped down to their simplest theoretical forms, where in EXP.2 further potentially predictive information was included in the modeling process. It seems likely that more extensive models (along with the required increase in participant numbers) may have found similar effects to those described in the EXP.2 analysis. Specifically, it seems possible that the inclusion of BinaryHours in the EXP.1 analyses could have found a similar lack of convergence/divergence for musicians who listen to little music daily, had participant numbers also been sufficient.

A reasonable question at this point would be: *Now that we've recognized these effects, why might they occur?* In other words, what do speakers gain by these changes in speech – or, at least what might be driving them to change in such ways? Because data from the pitch manipulation and much of the intensity data indicate that speakers who listen to relatively more music exhibit convergent behaviour and those who listen to relatively less are largely found to diverge, these results can be interpreted as support for traditional *socially-based* effects – specifically, in previous sociolinguistic works speakers are well known to converge with things they 'like' and diverge from things they 'do not like, or like less.' Inferring from speakers' listening habits, it seems logical that people who listen to less music may regard music less fondly, and those who listen to relatively more music may hold more positive opinions of music. Presuming this to be the case, the BinaryHours effects for the Pitch and much of the intensity analyses appear to fit well with previous research. Motivations become less clear, however, when considering the musicians and SMs who listen to relatively little music. Given the continuous nature of the SM category, it seems possible that the SMs in EXP.2 may have been closer to the musician side of the categorical cut-offs and were therefore performing more like the musicians in EXP.1. These speakers appear to have maximized differences between their speech signals and the background music, either in the name of intelligibility or by way of more socially driven motivations. If these SMs do not listen to much music, and therefore can be assumed to be less fond of music generally, then these patterns very much fit with a similar social drive for divergence. But what of the musicians who showed no effect? While speculative, perhaps these musicians who listen to relatively little music did not become musicians by choice. In this study anyone with six or more years formal training was classified as a musician; while not testable given the available data, it could be the case that the musicians showing no effect were enrolled in musical studies at a young age but did not particularly enjoy them. It could be the case that this group of musicians who listen to little music is comprised largely of somewhat *unwilling*-musicians, people who studied regularly for

years and lost interest and/or people who study/studied music for reasons other than a love for music (e.g., cognitive development, forced studies in *high culture* at a young age, etc.). Without qualitative data to confirm or disconfirm this hypothesis it remains speculation, though distaste for music could very reasonably account for both divergence and negated effects. This process would also fit well with a socially driven account of convergence and divergence in speech production.

It is also worth discussing the effects found above related to a speaker's identified gender. Recall that when the data were subset, male speakers were predicted to diverge and speakers identifying as female trended toward convergence (though the latter difference was non-significant). Numerous papers exploring convergence in speech production have noted complex and at times uninterpretable gender-based effects (e.g., Hay, Nolan & Drager, 2006; Hay & Drager, 2010), and Drager, Hay & Walker (2010) specifically describe female productions shifting in the predicted direction while male productions were observed to shift in the opposite direction. Given the implied complex interaction in the present analysis between BinaryHours, IdentGender, and previous musical training, and the limited representation across possible cells as models becoming increasingly complex, it seems entirely possible that exploration of a potential three-way interaction (in conjunction with an even larger dataset) might help better understand these effects. Analyses at this point do suggest, however, that these three variables are contributing individually, and at times together through some dynamic/complex relationship. Indeed, perhaps all three variables are also related to some unknown predictor that is actually the primary impetus for convergence in speech production to ambient noise.

With regard to methods, it is worth noting that changes made to avoid the analytical issues encountered in EXP.1 appear to have been highly effective. While the gradual manipulations over time employed in EXP.1 were theoretically motivated, they resulted in troublesome and unstable analyses. Altering the experimental design to instead involve global manipulations, where acoustic information was extracted as mean values by item, appears to have solved all methodological issues raised by this shortcoming. With regard to prosody and speech materials, performance within conditions appears to have been relatively stable – phrases read aloud in EXP.2 were not connected as any greater narrative, and for this reason appear to have not influenced talkers' speech patterns in unintended ways. It is possible that speakers may have exhibited list intonation to some degree, though this potential effect should not be any more or less present in any specific condition given identical presentation and randomization of stimuli across treatments, so has therefore been controlled as a result. Indeed, the increased participant numbers have also resulted in much more stable, and powerful analyses. For these reasons, the improved design will be largely retained for the next study.

While, ideally, models trained in EXP.2's post-hoc analyses would have been tested on data from EXP.1, much like the models from EXP.1 were initially used to test data collected in EXP.2, the nature of the time series data and low participant numbers in the first experiment would once again necessarily result in a low-power analysis. As a result, Experiment 3 (described in chapter 4) has been designed to directly address the issue of replicability and further exploration of the social and experiential predictors. This extension of the present methodology should allow for more confident conclusions regarding predictability and convergent tendencies.

At this point though, we know very little about how similar or different the processes driving entrainment to background speech and background noise may be. There have been no studies directly comparing these processes, where such a comparison may speak directly to the roles of social and experiential components in acoustic convergence during speech production. Moreover, only work from Delvaux & Soquet (2007a, 2007b) has thus far approached convergence to ambient speech, and a replication exploring similar forms of convergence would be of theoretical value, as the work described by these authors involves certain design-based issues that also drive alternate interpretations of their findings.

Specifically, Delvaux and Soquet (2007a; 2007b) claim that presentation of stimuli in their work approximates passive exposure to ambient speech; however, I am reluctant to accept this as the case. In fact, given descriptions of their methods, I would argue their task better simulates situations that are more communicative and interactive in nature. Recapping their description, participants sat at a table with two other *virtual* participants i.e., loudspeakers on either side of that table which reproduced the appropriate prerecorded speech. Participants were directed to speak aloud when their "turn came" (i.e., one third of the stimuli, where the other 2/3 were 'spoken' by the virtual participants). In fact, in their participant-debriefing Delvaux and Soquet describe a line of questioning involving the other participant "who was hiding" (2007a). Considering the task is effectively an exercise in *turn-taking*, as well as the fact that the experimenters chose to reinforce the notion of speakers alternating with other participants (even if after the fact), I would argue this work does not adequately replicate passive exposure to ambient speech. Instead this experiment seems much more like experiencing spoken language without the aid of visual cues, much like a telephone conversation. I believe participants would be much more likely to attend directly to these speakers and not disregard them as noise in the context described by these authors, despite the experimenters' instruction to actively ignore the virtual speakers. By this reasoning I believe another experimental paradigm would better simulate convergence to ambient speech.

Acoustic convergence to ambient speech could, perhaps more accurately be simulated by prompting participants to produce speech while experiencing a relatively constant stream of speech or conversation, which would eliminate the communicative context of a task centered around turn-taking. The proposed exercise could be made stronger yet through use of headphones and a mono-mix instead of allotting positions at the table for virtual speakers, where the speaker-separation reinforces perception of distinct entities at either side of the table. Two possible forms of background speech-noise could be effective, where the most appropriate form should be selected given the hypotheses and experimental context: (1) A single speaker could be recorded and filtered in a way that reproduces the experience of hearing only half of a telephone conversation with the audible speaker in an adjacent room, or (2) Two or more speakers (matched for dialect and controlled for sex) could be recorded in conversation and presented to the participant as if in the same room, which would be very much like incidentally overhearing a conversation at another table within a restaurant or other public space. In the context of the present work, I have opted for the latter to more closely resemble musical stimuli, and to avoid confounds related to gender-biased convergence.

Firstly, the musical stimuli (Science Music) is comprised of multiple instrumental voices, any of which may be 'latched onto' by a participant for one reason or another as the home frequency. Given findings from previous

work, which component-voice is 'selected' by a participant is not necessarily expected to be consistent across talkers (Pardo, 2013), which is why the present analyses explore relative changes in performance (as opposed to absolute). Therefore, presenting multiple voices in a speech-based condition more closely approximates musical stimuli than would a single speaker's voice. Secondly, we know from previous work that speakers are more likely to converge with speakers they identify with or hold positive opinions of (Giles, Coupland & Coupland, 1991; Drager, Hay, & Walker, 2010; Giles & Powesland, 1975; Babel, 2012), which could well be reflected in a gender bias. With such potential biases in mind, it seems good practice to present subjects with recordings comprised of *at least* one female and one male speaker. Methods of EXP.3 will be discussed in more detail in the following chapter.

3.12 Conclusion

Experiment 2 primarily builds upon Experiment 1 through improved methodology with regard to both experimental design and the approach to data analysis, improved participant numbers, and results that aid in better understanding the nature of acoustic convergence to (and divergence from) ambient noise presented as background music. These results appear to harmonize well enough with those of Experiment 1. Where EXP.1 provided evidence which could be interpreted as supporting acoustic entrainment to background noise – though, could also be explained through other processes – EXP.2 provides further evidence (through a more robust analysis) that speakers are influenced by background noise, and that convergent and divergent behaviours were both individually significant. Given the preceding analyses, these processes appear very much like socially driven convergence and divergence. Results from Experiments 1 + 2 also support the possibility, however, that the variation encountered in speech produced against background noise can be predicted somewhat through speakers' previous experience.

Simply, it seems variation encountered in speech across productions in this study is not random. Specific variables have been identified by acoustic manipulation, and some variables were shared across conditions which have been shown to predict speaker behaviours reasonably well. Of those predictors, *previous musical training* may not generally be as predictive in this context as once thought, or at least not always, but was still shown to play an important role. While the random forests in EXP.2 found musicianship to be useful when predicting speaker intensity, the role of previous musical training was found to be reduced somewhat when compared to that in EXP.1. Importantly, musicianship was not involved when predicting behaviour during the Pitch treatment in these data. The random forests and mixed effects models together, however, support an important contribution from HoursMusic (that is, the estimated amount of music a speaker listens to daily), which would be expected to correlate with Musicianship to some extent – and it does, though in a counter-intuitive way: Across EXPs 1 + 2 musicians often self-report as listening to less music per day than the non-musicians. In the context of ambient noise then, data suggest that musical training, the HoursMusic variable, and a participant's identified gender are all important to some degree when considering convergent/divergent behaviour in speech production in this experimental context. It has been hypothesized, though, within the present work, that they may be involved together in a dynamic relationship or are perhaps related to some unknown predictor, which could in fact be the primary drive for these altered speech productions.

While different approaches to analyses in EXP.1 often returned conflicting results, random forests and mixed effects models in the present work were found to provide complementary outputs for the most part, and provided results that largely fit well with the final analysis described in EXP.1 (keeping in mind the models in EXP.2 were markedly more complex). Surely, dealing with the autocorrelation issues through an altered experimental design in EXP.2 played a major role in identifying trends in the data more accurately, and it is encouraging to have found statistical methodologies that appear more robust while also telling a similar story.

In the way of results, we have gathered further support that participants do, in fact, seem to converge and diverge from both pitch and intensity variation encountered in ambient music, and that these differences in response type can be predicted to some extent by music-related experience (playing and/or listening): That is, speakers who listen to relatively more music (are assumed to be fond of music for this reason and) are found to converge generally, while speakers who listen to relatively less music (are assumed to be less fond of music and) have been observed generally to exhibit divergent behaviour. Indeed, data suggest that some of these intensity-based effects are further mediated by a speaker's identified gender. We have also found that speakers do appear to be influenced by intensity envelopes below the known threshold for eliciting Lombard speech; this latter finding is important insofar as it shows either that (1) Lombard responses and entrainment effects are distinct events/processes, or (2) that the known threshold for eliciting the Lombard effect is, in fact, lower than previously believed. Crucially, many of the effects observed in this work can be explained in a straightforward way by effects observed in previous socio-phonetic research (i.e., alignment and speaker attitudes), suggesting entrainment and the Lombard effect are likely distinct processes.

In summary, this work has supported acoustic convergence and divergence to background music in speech production, and the specific effects observed are largely explainable through previous accounts of accommodation observed in communicative speech. This study, therefore, plays an important role in better understanding both auditory processing and human cognition more generally – where results provide insight into how hearers/speakers deal with various noise types in processing, and how we can be influenced in different ways leading up to production. As a result, it seems likely that this work could also feed the development of theories addressing the complex relationship between speech production and perception. One way to further this line of research would be to once again explore the notion of entrainment to *speech*-as-background-noise in order to ensure previous descriptions of similar effects were not in fact artifacts of experimental design. Moreover, a general understanding of convergence in speech would benefit greatly from an experiment that allows for direct comparison of entrainment to speech vs. entrainment to non-speech noise, where the magnitudes of these effects – being similar or different in any reliable way – would speak directly to arguments regarding the cognitive mechanisms involved in auditory processing (e.g., Liberman, 1984). The study that follows in Chapter 4 includes both musical and speech-based stimuli, and was designed specifically for such direct comparison.

CHAPTER 4: Experiment 3 (Further Replication, and the Addition of Background Speech)

4.1 Introduction

When considered together, results from Experiments 1 & 2 provide reasonably compelling evidence for a predictable influence of ambient music on the fine phonetic detail in human speech production; however, the specific drive for such convergent/divergent behaviours is still not completely clear. We have seen that a speaker's musical experience, musical listening habits, and a speaker's identified gender all seem to play important roles when predicting certain forms of acoustic convergence & divergence. Thus far, self-reported estimates of 'music listened to per day on average' appear to provide the most robust explanatory power; specifically, it appears that speakers who listen to relatively little music per day have been diverging, whereas speakers who listen relatively more are converging with acoustic characteristics of background music. These effects appear, though, to be mediated somewhat by a speaker's musical training.

We have also seen that certain forms of pitch-variation appear to be reflected in speech production for both rising and falling trajectories in the stimuli (as a function of convergence or divergence), and that this was also the case for intensity-based variation in both of the preceding experiments. However, while this finding is interesting in the case of entrainment to pitch variation, it is particularly remarkable in the case of convergence to intensity, where the latter process was observed in EXP.2 below a presentation level of ~45 dB(A), the known floor for eliciting the Lombard effect (Lombard, 1911; Lazarus, 1986). Therefore, not only have EXPs 1 + 2 shown that background noise (in this context, background music) can reliably influence speech production in ways that are predictable using speakers' musical consumption habits, but also that speakers are sensitive to intensity-based variation *at least* 6 dB below the level previously believed. Of course this result does not conclusively distinguish convergence to intensity as distinct from the Lombard effect, at least without further spectral analyses of the data; though, it does provide evidence that either (1) Convergence to intensity and The Lombard effect are distinct processes, or (2) that the known floor for eliciting the Lombard effect is incorrect, and is in fact lower than previously recognized. It has been recognized, however, that many of the effects observed in the previous two studies are readily explainable through comparisons to previous socio-phonetic works exploring convergence, suggesting a motivation likely rooted in alliance and speaker attitudes, and therefore a distinction between entrainment and the Lombard effect.

The focus of this dissertation is on acoustic convergence to ambient noise though, and not the Lombard effect. Therefore, having provided reasonable evidence that speakers appear to converge acoustically with background noise both above and below ~45 dB(A), the experimental focus in EXP.3 is restricted to acoustic/phonetic-convergence to pitch only. Studies 1 + 2 have contributed to better understanding the influence of ambient non-speech noise on a speaker's productions, but the work in this collection thus far have relied wholly upon descriptions from only two pre-existing, related studies with regard to the possibility of convergence to ambient

speech-noise, both employing an arguably flawed methodology (Delvaux & Soquet, 2007a; 2007b). Moreover, an understanding of convergence to ambient sound generally would benefit from a direct comparison of entrainment-effects involving background speech with entrainment-effects involving background noise. Experiment 3 introduces new pitch-based treatments as a result to more effectively explore the potential for convergence to voice-pitch in ambient speech, in addition to those treatments already exploring comparable effects in background music. Crucially, the present experiment has been designed with an analysis in mind that allows for direct comparison of any effects that may be observed across conditions.

The works from Delvaux and Soquet (2007a, 2007b), involve some notable design choices that are likely to have influenced participants in unintended ways, specifically leading listeners to regard "background" noise as less ambient and more *interactive* in the experimental context. As mentioned in the previous discussion (Section 3.11), Delvaux and Soquet claim that participants experience conditions approximating passive exposure to ambient speech, though descriptions of their methods suggest laboratory situations which are more communicative and interactive for participants. Their subjects were seated at a table in between two loudspeakers, where a computer monitor would indicate whether speech was meant to be produced by loudspeaker #1, the participant, or loudspeaker #2 – which, in practice, is much more an exercise in turn-taking than in passive exposure. This design choice seems a real issue in this context, given the large body of previous work crediting social contributions to various entrainment effects (described throughout this dissertation). Of course, some degree of passive exposure must surely also be experienced by participants in these experiments, though it seems likely that this kind of exposure would carry less weight than the implied participation of other speakers through the loudspeakers and seating arrangements. When 2/3 of the stimuli within a session were produced by virtual participants, it seems probable that experimental subjects might feel as though they were interacting with the talkers described by Delvaux and Soquet as "hiding". Given this confound, their work may not directly test for convergence effects related to ambient speech (as opposed to communicative speech), and I believe a redesigned condition may better explore any such potential influence. Therefore, Experiment 3 (described below) includes test conditions related to both ambient speech *and* music, designed for direct comparison within and across conditions.

Note though, that specific predictions are difficult to make at this point, because the most robust predictor leading up to EXP.3 has involved speakers' musical consumption in the context of entrainment to background music. There is no reason to expect that speakers entraining to ambient speech should be influenced similarly by way of such consumption, and no directly comparable data have been collected which might be tested for potentially influencing convergence/divergence to speech in a similar way – though, the potential for a relationship involving musical consumption is tested explicitly.

4.2 Overview and Generation of Background Stimuli

The same musical baseline and lowered pitch background stimuli used in EXP.2 were also used in EXP.3, and these will be referred to below as the *music-based* stimuli This choice was made to maximize comparability across studies,

and to retain equipment calibration with regard to the perceived loudness of background stimuli when presented to participants. Experiment 3, however, introduces two new conditions in order to more effectively test for the possibility of entrainment to ambient speech, and to explore potential differences across background speech and music: Specifically, these new conditions include (1) a baseline background speech condition, and (2) a lowered-pitch (manipulated) background speech condition – both comprised of multi-talker babble. The new treatments will be referred to below as *speech*-based stimuli, and were generated as follows.

Two native speakers of New Zealand English, one male and one female, were recorded speaking in the same sound attenuated booth described in EXPs 1+2. Both speakers were 23 years old at the time of recording, and were born and raised in the Hutt Valley area of Wellington, NZ. The decision to have one male and one female speaker was made to avoid potential gender-based biases in convergence (Namy et al., 2002). These two particular talkers were selected for being close friends with the aim of capturing a reasonable approximation of casual speech (much like the methodology described in Warner, 2012). Each speaker was equipped with a Beyerdynamic Opus 55.18 MK II head mounted condenser microphone, and each mic was routed through a separate channel in a Sound Devices USBPre 2 audio interface. Segregating microphones in this way facilitated higher quality manipulations by allowing for adjustment of the pitch and intensity for each speaker individually. Using Praat (Boersma & Weenink, 2014), speech was recorded as .wav files on the same Macbook Pro laptop computer described in experiments 1 & 2 at a sampling-rate of 44.1 kHz and 16 bits. Three clips of 5 minutes 40 seconds were recorded; while some subject-matter from the end of one clip could bleed over into the next, topics generally varied widely over the course of each recording as a conversation progressed.

One clip was selected on an impressionistic basis for containing the least amount of laughing, coughing, and pauses for use in the final stimuli – this clip will be referred to as *Conversation 1*. This process was repeated with a married couple also comprised of one male and one female native speaker of New Zealand English, aged 43 and 39 years respectively, both of whom were born and raised in Christchurch, NZ. The clip selected from the second couple will be referred to as *Conversation 2*.

Speech-based stimuli must be matched for duration to Science music. To this end, selections of 207 seconds were extracted from each longer conversation (placing boundaries at zero-crossings to avoid unwanted pops at signal onsets and offsets). For simplicity's sake, these selections are still referred to below as 'conversations' 1 + 2. When generating these kinds of signals, it would be prudent to ensure speaker-loudness has been matched across conversations. Using Praat, the left and right channels were next extracted from each clip. In other words, speaker 1 and speaker 2 were extracted as independent sub-recordings from each conversation. Though speaker-intensity was approximately matched at the time of recording using an LED VU meter, this method lacks the precision of directly scaling each signal to a designated value. Mean intensity was therefore extracted for each speaker and the highest intensity value from the four speakers (66.909 dB) was then selected as a target for scaling all other recordings.

Here, it would be worth mentioning a concern rooted in priming and attention recognized in the early stages of stimuli development. Initially the two conversations (each comprised of two unique tracks/speakers) were combined as a single track at this point in the process – that is, layered on top of each other – in order to (1) have multiple voices sounding simultaneously, making this treatment more comparable to the music condition, and (2) to

make the speech less intelligible and, thus, less of a distraction/draw for attention to listeners during the study. However, specific words and phrases began to 'pop out' of the babble as relatively more salient to pilot listeners. This recognition seems not unlike what Cooke (2006) describes as phonetic *glimpses*. While describing speech in noise processing, the author notes that:

> "…since speech is a highly modulated signal in time and frequency, regions of high energy are typically sparsely distributed. Consequently, the spectrotemporal distribution of energy in a mixture contains regions that are dominated by the target speech source, even at quite adverse signal-to-noise ratios."

Put simply, when scaling by average intensity, the speech produced will still exhibit a range of high and low levels, and will therefore at times be subject to relatively increased and decreased levels of masking as a result of within-signal intensity-variation. The potential for glimpses is complicated further by the balancing of multiple concurrent speakers, where others may in fact be at relative low-points while another speaker hits a relatively louder moment. Therefore, where the two speech-based background stimuli (i.e., baseline vs. lowered pitch) were initially designed to be identical in every way with the exception of lowered pitch in the test condition, as is the case in the musical analogs of these treatments, this plan could cause new problems in this context. While for many listeners it is possible for a musical piece to be performed multiple times in different keys and not be recognized necessarily as a single, manipulated performance (where the relative movement and voicings of each instrument are precisely the same, though transposed), listeners reliably recognized certain elements and topics of speech as repeated in these stimuli, even with augmented pitch. If participants are recognizing the two speech-based conditions as effectively the same signals, the constructed nature of the task becomes rather glaring. Therefore, in order to retain the majority of the acoustic content across speech-based treatments (i.e., comparability across conditions), while also aiming to (1) alter which segments pop out as more salient and, in turn, (2) give the impression of non-repeated speech across conditions, the sub-selections from conversations 1 + 2 have been slightly shifted when generating the test stimulus. Specifically, 80% of conversation 1 and 80% of conversation 2 are included within the lowered-pitch speech stimulus, though 20% of each conversation is now new to speakers, and how the conversations align to obscure each other has also been shifted. A mock-up is available as Figure 4.1 (below), approximating how these selections have been processed. I will refer to conversations 1a and 2a when referencing the initially extracted segments, and 1b and 2b when referring to their temporally shifted variants.

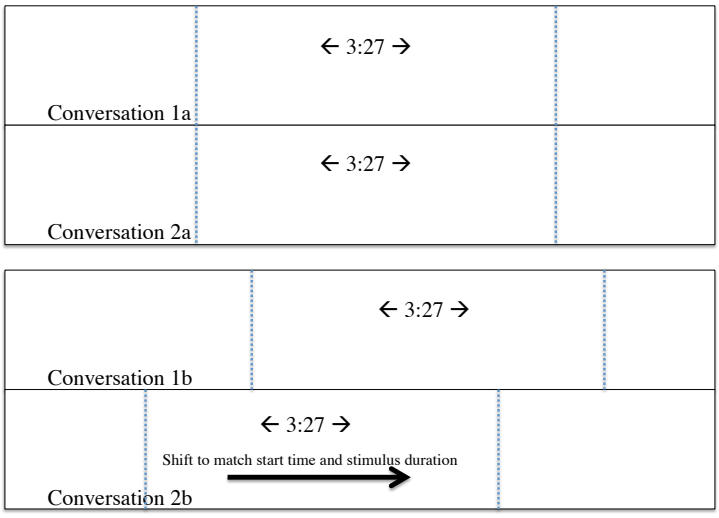

Figure 4.1 This diagram illustrates how sub-sections of conversations 1 and 2 were selected and combined. Three minutes and twenty-seven seconds were selected from each clip and overlaid to create the babble-like stimuli used as the baseline condition. The time points where each sub-section begins were shifted for each conversation (in opposite directions) to generate the test stimulus prior to pitch manipulation, maintaining 80% of the acoustic content from each sub-section while varying the dynamic interplay of how the conversations mix together as babble. Note: components in the "shifted" stimulus were aligned to begin at the same time, maintaining the desired 3:27 duration for both speech-based stimuli.

Next, the global voice-pitch of each speaker in conversations 1b and 2b was lowered by 2 semitones, or 200 cents, to maximize comparability with musical stimuli in the Pitch manipulations described in EXP.1 and EXP.2. The process for this manipulation is available as Appendix 5. Now unaltered and pitch-altered speech-based sound files exist insofar as both have been matched for within-signal speaker intensity, and for global duration to Science Music (that is, 207 seconds). Mean Pitch was extracted from both sound-files as a precautionary measure to ensure manipulations were applied as intended, confirming the pitch-altered stimulus was indeed composed of lower F0 values (MeanPitch UnAltered: 143Hz; MeanPitch Altered: 130.6 Hz).

One final issue that must be addressed before the speech stimuli can be scaled to the mean intensity of Science Music is that of *compression* (for detailed discussion, see EXP.1 – *Specific Considerations During Stimuli Manipulation: Intensity* above). Science Music has been compressed to maintain a relatively static intensity level, where the intensity levels of speech in these new recordings are considerably less stable. Therefore, the new speech-recordings were also compressed as described earlier (using Ableton Live 9, 2015) before finally being intensity-matched to Science music in Praat. While it is true that Praat can easily calculate averaged perceived loudness across files using RMS amplitude only, the use of a compressor results in averaged intensities that are more directly comparable; the reason for this difference lies primarily in deviations from the mean intensity value. Much like the above mention of "glimpsing", it is possible to choose identical mean values that result in relatively higher highs and lower lows across files when scaling based on RMS amplitude (in fact, in speech recordings you would often want to

retain this prosodic information). Compression in Ableton, however, results in much more consistent intensity-levels, which can be seen clearly through the visualizations of the stimuli pre- and post-compression stimuli, available below in figures 4.2 and 4.3.

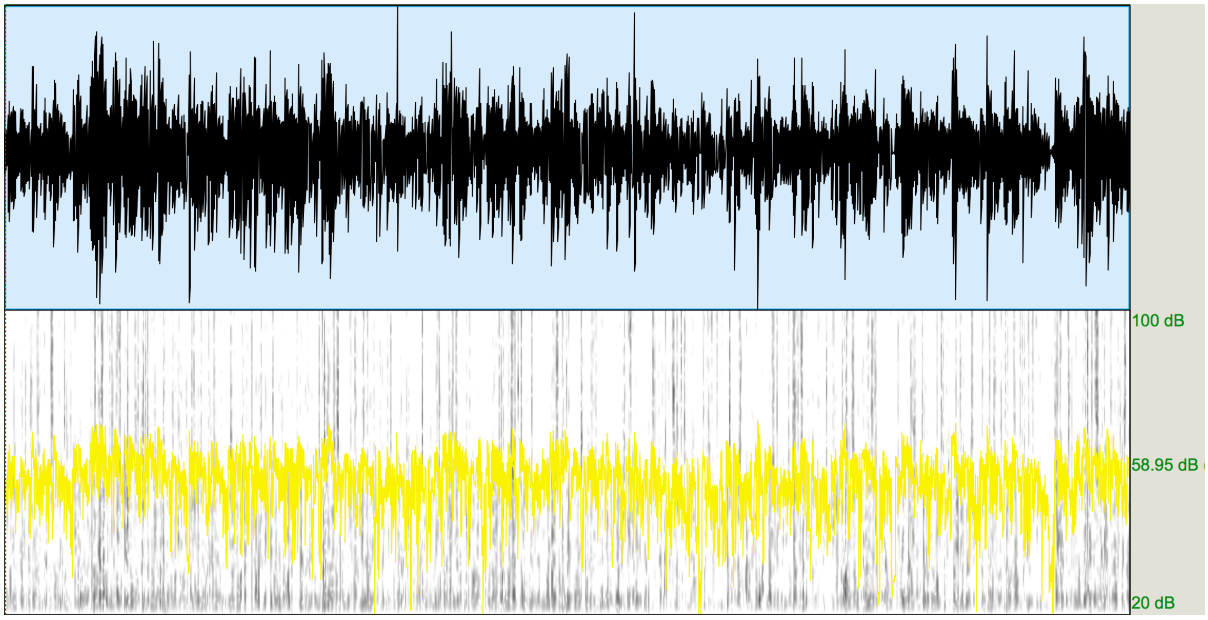

Figure 4.2 A screen-grab from Praat showing the waveform and intensity contour of the speech/babble before undergoing compression.

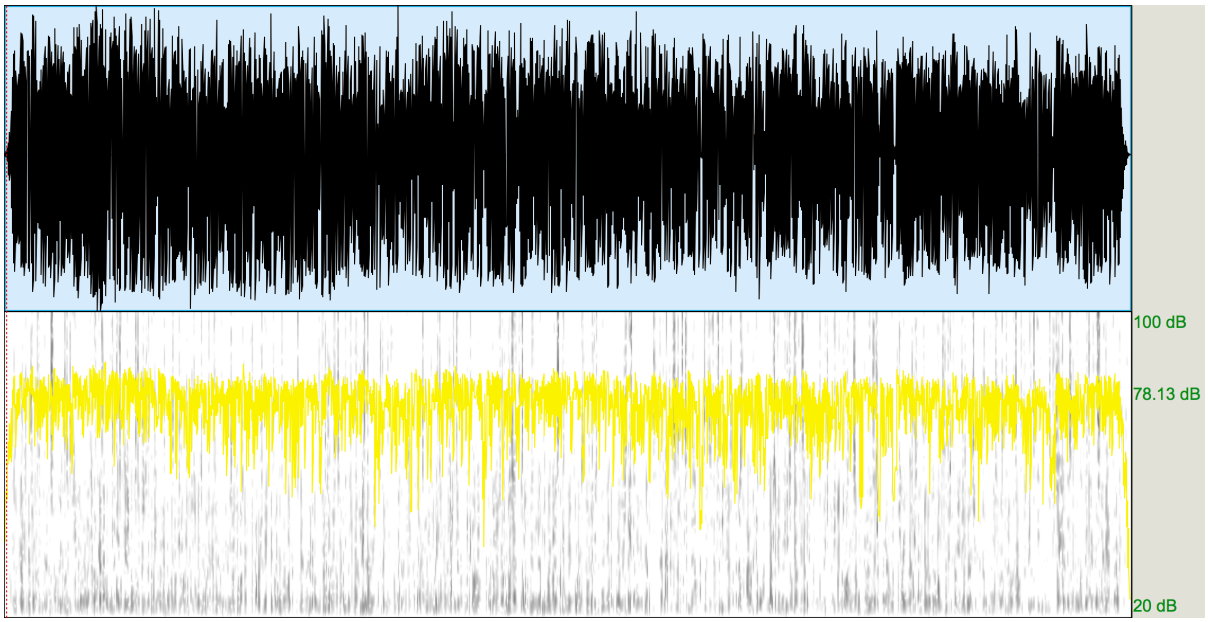

Figure 4.3 A screen-grab from Praat showing the waveform and intensity contour of the speech/babble after undergoing compression. Note the leveling of peaks and troughs in both the waveform and the intensity envelope, as well as the additional boosted level that can come as a result of such leveling (without distortions due to digital clipping).

Found within Ableton's onboard compressors, "The Flatenator" preset was used to compress the signals as above. Good results were achieved with only minor adjustment (Ratio: 4.67:1, Attack, 0.01ms, Release: 113 ms, Dry/Wet: 100%). Now that the majority of variation in the intensity envelope has been removed, these signals can be scaled to match the music-based stimuli for perceived loudness. Mean intensity was therefore extracted from the Baseline version of Science Music by selecting it as a 'Sound object' in Praat, and using the "Query" button to access the *Get intensity (dB)* function. This intensity value (74.4853 dB) was then used as a target for the new speech recordings to ensure a more precise match.

Along with the previous Baseline and Pitch-altered versions of Science Music, all speech-based background stimuli for EXP.3 have also been fully prepared at this point. These stimuli are available for download through the following link: https://github.com/RyanPodlubny/ScienceMusic_EXP3

4.3 Production Stimuli

Production stimuli in EXP.3 were identical to those used in EXP.2, with the following exception. Adding a fourth block to the study meant that either fewer stimuli would be presented in each treatment, or that 54 new production stimuli must be generated under the same constraints as the original production stimuli set (i.e., comprised of all voiced sonorants). Additional production stimuli were therefore generated to retain maximal comparability across studies and conditions, and are available as Appendix 6.

4.4 Equipment and Stimuli Calibration

The equipment used in EXP.3 was identical to that used in EXP.2. All hardware and settings were maintained in order to retain the desired presentation volume of stimuli (i.e., ~45 db(A)). New stimuli were scaled as described above (that is, scaled to match levels from older stimuli) to ensure the desired presentation levels were consistent across both treatments and studies.

4.5 Participants

40 native speakers of New Zealand English took part in this study. All speakers were recruited via posters displayed on campus, adverts through various forms of social media, and brief discussions in entry-level linguistics and music courses. Participants received a $15 NZD voucher for use at a local shopping centre in exchange for their time. The group was comprised of 17 males and 23 females (no participants reported identifying as non-binary); no efforts were made to counterbalance for degree of musicianship in this study because listening experience appears to be more important, and recruitment becomes much easier without trying to balance cells for musicianship. Seven of

these speakers (3 male, 4 female) were excluded from analysis due to equipment failure during data collection, or because they had grown up speaking a language other than English at home. As a result, data from 33 speakers in total (11 Musicians, 14 SomeMusic, 8 Non-musicians; 19 Female, 14 Male) are analyzed and described below.

4.6 Procedure/Protocol

The procedure during EXP.3 was nearly identical to that of EXP.2. Upon arrival subjects were assigned a participant number encoding meta-information formatted exactly as in EXP. 1 + 2 (that is, including the experiment name, the year of participation, the participant's chronological rank in the study, identified gender, and musical training category). Subjects were asked to read through and complete the same consent form and language background survey used in the previous studies, and sat a full hearing screening prior to beginning the experiment. An Interacoustics AS608 audiometer was used for all screenings.

Each session took place within the same sound attenuated booth described in EXP. 1 + 2, and participants were outfitted with the same equipment used in EXP.2 (i.e., a Sound Devices USBPre 2 audio interface, a Beyerdynamic Opus 55.18 MK II head-mounted condenser microphone, and a pair of Audio-technica ATH-M40x closed back, isolation headphones – all of which were routed through E-Prime 2.0 installed on a Hewlett Packard EliteBook 850 G3 laptop computer). Participants encountered four treatments (being BaselineMusic, LoweredPitchMusic, BaselineSpeech, LoweredPitchSpeech) in a given session, where the presentation order of these conditions had been counterbalanced. All treatments were once again broken up by an unrelated task (i.e., the same video game used in EXP.2). Production stimuli were presented visually on a computer monitor for participants to speak aloud in each condition; all production stimuli remained onscreen for 3.5 seconds and were replaced by a blank screen before presentation of the following stimulus. Once again, the duration of the blank screen was a randomized value between 250 and 500 milliseconds. Presentation of the production stimuli continued in this way for each background stimulus. Production stimuli were selected randomly without replacement from the list, and all items were encountered once (and only once) in a given session. Background stimuli were again presented at ~45 dB(A) as in EXP. 1 + 2. In between each condition participants were instructed to shift to the adjacent station/laptop in order to complete one level in the video game in silence. As in EXP.2, one level of "Quinn" involved playing either until the participant had scored 10 lines, or had unintentionally ended their turn by stacking game pieces to the top of the screen. Signage was placed in between the two computer stations to remind participants how to navigate the different conditions. All participants completed the same debriefing survey used in EXP. 1 + 2 following the fourth test condition.

4.7 Analysis

The analytical methods used in the following discussion are identical to those described in EXP.2. All predictor variables and their corresponding labels (described in section 2.8.2) have been carried over from the previous two

studies. The following analyses explore convergence in voice-pitch to both ambient music and ambient speech. Once again, mean F0 was extracted by item (i.e., for each production stimulus), resulting in the collection of 7,127 observations. Predictions are unavailable for the speech-based condition, though it is predicted that convergence and divergence to background music will once again be influenced by speakers' music-listening habits (i.e., more music = convergence; less music = divergence). As a first step, the procedure outlined in EXP.2 (section 3.10.1) was repeated in order to generate t-values for each participant by condition. These values reflect a normalized measure of effect size and direction in treatment conditions as distanced from their corresponding baselines, where the baseline for each speaker/condition is treated as the 'zero' mark. This effort serves largely as a descriptive, exploratory tool allowing for any clear trends to be recognized more easily. These scores, once again referred to as *t-values*, reduced the dataset to 66 observations that will be explored through random forests, and have also been plotted below as Figure 4.4.

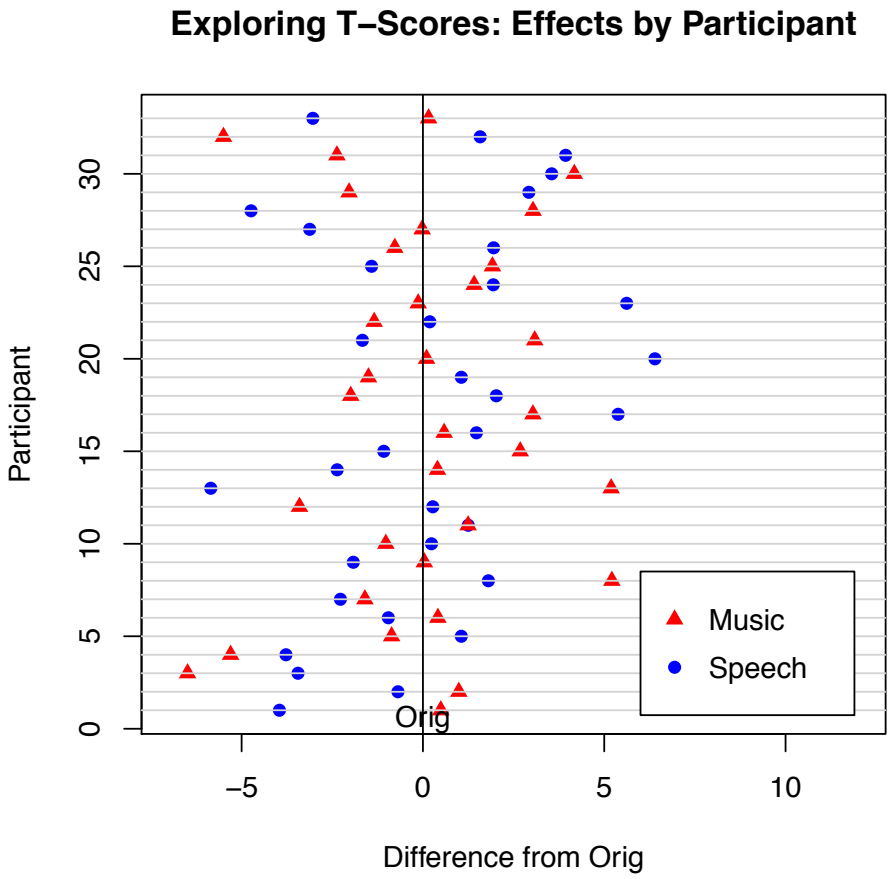

Figure 4.4 This figure expresses effect size and direction by participant, by condition. Participant number increases along the y-axis, where normalized effect size can be seen on the x-axis. Red triangles indicate t-values for the Musical treatments and blue circles indicate those of the speech-based conditions; the "Orig" divider line is treated as zero, and was derived as a relative distance from treatment t-values by condition for each participant.

As was the case in EXP.2, no clear patterns were immediately visible through this figure. Correlations across conditions are once again tested below through predicted data (Figure 4.6) as well as more directly; though, it should be noted that any speech-based vs. music-based effect comparisons and correlations are not subject to the same confounds observed in EXP.2, as the baseline conditions used to calculate this transformation were not shared across conditions. Further note that at the ≥2 criterion, the majority of these effects would be considered statistically significant. Once again we find participants converging in both treatments (n = 4); diverging in both treatments (n = 7); and some showing inconsistencies in effect-direction across conditions (n = 22). Of the twenty-two participants who converged in one treatment and diverged in the other, exactly half exhibited convergence only to background speech, while the other half exhibited convergence only to background music.

Contrasting the above descriptive analysis, all remaining analyses utilize more inferential statistical methods. The analyses that follow were conducted in R (R Core Team, 2013) using the lme4, languageR, lattice and Party convenience packages (Bates, Maechler, Bolker, & Walker, 2015; Baayen, 2013; Sarkar, 2017; Hothorn, Buehlmann, Dudoit, Molinaro, & van DerLaan, 2006; Strobl, Boulesteix, Zeileis, & Hothorn, 2007; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). Much like the Post-hoc analyses described in EXP.2, I will first discuss the results of some random forests before outlining how those results were used to inform the subsequent linear mixed effects models. All forests were fitted to the reduced t-values dataset (66 observations), while the mixed effects models were fit to raw data (7,127 observations).

First, because there is no reason to expect speakers will respond uniformly across speech-based and music-based treatments, data were split to model observations from each separately. This sub-setting resulted in 33 observations per forest. Forests were grown predicting t-values, and included Participant, Condition (Treatment vs. Baseline), Age, IdentifiedGender, Handedness, Musicianship, HoursMusicPerDay (self-reported estimates of hours spent listening to music daily), and ChooseMusicWhileWork (whether or not the speaker chooses to listen to listen to music while performing cognitively demanding tasks, henceforth *ChooseMusic*) as potential predictors. Mtry was set to "2" in all forests, and nTree was set to "5000". Please recall that random forests are visualized as Variable Importance Plots (VIPs), and that the absolute values reported in these plots are not meant to be interpreted. VIPs for each model are available below as Figure 4.5. We see that in the present data, convergence to background speech was best predicted by the listener's previous musical training, whereas convergence to background music was best predicted by whether or not the speaker chooses to listen to background music during cognitively demanding work tasks.

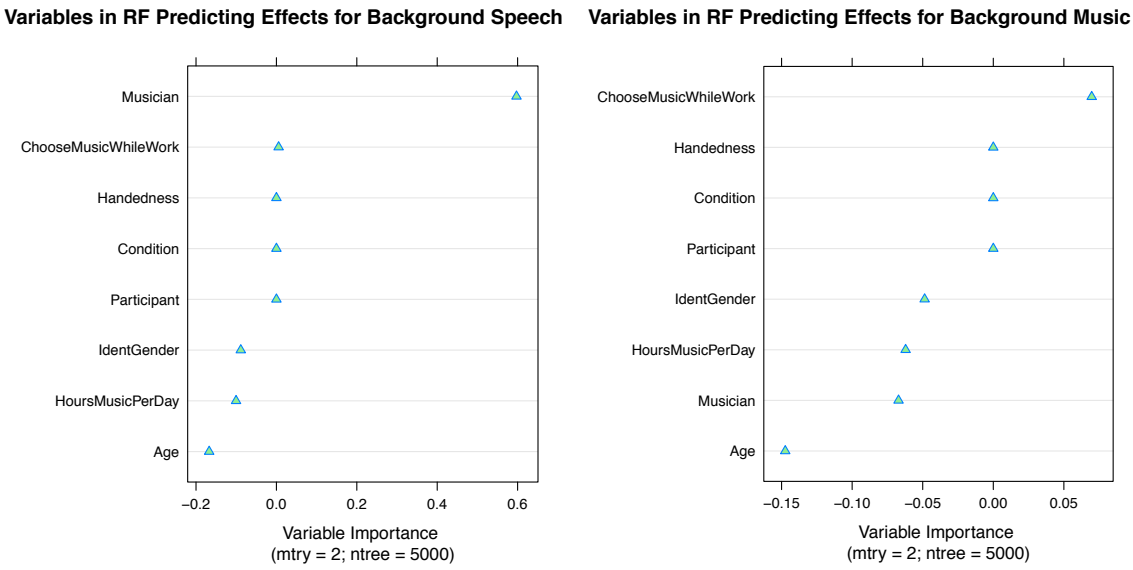

Figure 4.5 (Left) A Variable Importance Plot expressing which predictors were found to reliably influence convergence to ambient speech. In this case, only a speaker's previous musical training was useful when predicting convergence. (Right) A Variable Importance Plot expressing which predictors were found to reliably influence convergence to ambient music. In this case, only the ChooseMusic variable was found to be predictive of convergence.

Next, as in EXP.2, data were simulated using the random forest models fit for each condition (i.e., speech-based vs. music-based) to test whether or not the model for one condition could reliably predict performance in the other. Having generated the simulated data, these can then be tested for correlations against the known data from that condition. The following analyses serve two main purposes: (1) By using the model from one condition to predict (known) data for the other, they test for a participant's consistency in performance when comparing speech-based and music-based responses, and (2) By using models fit to t-values generated using non-related baseline conditions, this test breaks the link across conditions that was shared via t-values in EXP.2. Once again, generating these predicted data involved use of the *predict()* function in R, and simulated data were then tested for correlations with the *actual* data from that condition. Results of these correlation tests have been plotted below as Figure 4.6, along with the p-values for each test. These plots clearly show that, at least in these data, convergence/divergence to background music does not appear to be predictive of convergence/divergence to background speech. A more direct correlation test comparing t-values from the speech-based condition to those from the music-based condition further supports these observations as not related ($r = -0.0063$; $p = 0.9721$). These correlation tests may support the explanation provided in EXP.2 which argued correlations persisting after shuffling data were at least partially the product of how the t-values had been generated (i.e., a spurious correlation due to the shared baseline condition); though, it also remains possible that this relationship was partially driven by a link between pitch based and intensity based convergence to music, where no such link exists between convergence to music vs. convergence to speech. Considering the inherent physiologically driven connections between voice-pitch and vocal intensity, the latter account is not unreasonable.

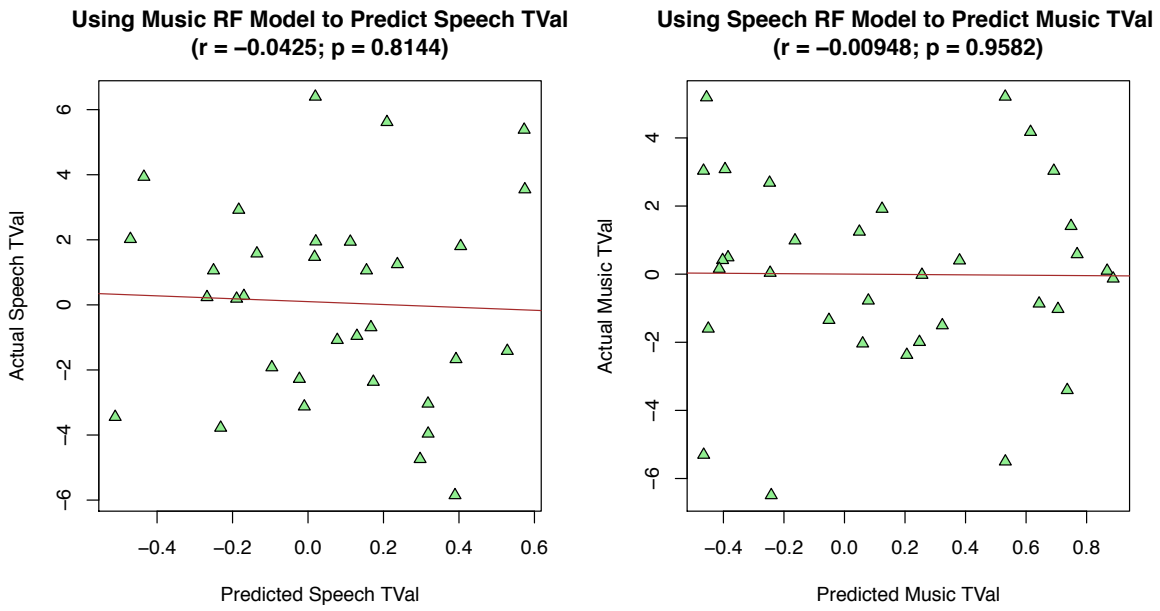

Figure 4.6 (Left) A scatterplot showing the correlation between actual data, and those predicted using the random forest modeling convergence to music. (Right). A scatterplot showing the correlation between actual data, and those predicted using the random forest modeling convergence to speech. Both plots show that convergence in one condition is not useful when predicting performance in the other.

4.7.1 Music-based Mixed Effects Models

Now that we have recognized certain predictors as important through the random forests, we can use mixed effects modelling to discover more about the specifics of those predictors. Most importantly, we can use the mixed models to learn about interactions. One major shortfall of random forests is the fact that they output only the importance of each variable, and give no information about how they may combine to influence situations. Using these statistical methods in tandem maximizes what we learn about each predictor in this context.

As in the above analyses, data explored in the mixed effects modelling were also subset by treatment (i.e., speech-based vs. music-based). Recall that the mixed effects models have been fit to raw data, unlike the transformed t-values used in the random forests. While the forests are well known for dealing with 'small n large p' problems, more data typically result in more robust mixed models. Therefore, from the 7,127 observations collected, 3,564 were analyzed when exploring responses to the music-based treatments, and 3,563 were analyzed in the speech-based analyses. Models described below were fit using a backward, stepwise process when testing for main effects. Subsequent models were compared via ANOVA tests in R, and those with lower AIC scores were preferred.

When testing for convergence in voice-pitch to ambient music, the ChooseMusic variable was included as suggested by the random forests; IdentGender and Condition were also included in the modelling process as control variables. All possible two-way interactions were explored. Random intercepts were included for Participant and Item, as were random slopes by Condition. The optimizer was set to "bobyqa". This final model is available below as Table 4.1.

Fixed effects:

	Estimate	Std. Error	t-value
(Intercept)	192.585	7.3927	26.051
ConditionOrigAltered	-0.3745	0.769	-0.487
IdentGenderM	**-79.4515**	**9.2865**	**-8.556**
ChooseMusicWhileWorkY	-5.0437	9.2175	-0.547
ConditionOrigAltered:IdentGenderM	**2.8554**	**0.9648**	**2.96**
ConditionOrigAltered:ChooseMusicWhileWorkY	**-2.6996**	**0.9574**	**-2.82**

Table 4.1 The final model predicting mean F0 in the Background Music condition, using predictors from the random forest modeling. Significant effects have been bolded for convenience.

The model indicates that, as would be expected, male speakers generally exhibit a lower F0 than do female speakers. We also see that male speakers tend to maintain significantly higher voice-pitch than the female speakers during the lowered pitch condition (Figure 4.7). Finally, this model indicates that in the lowered-pitch condition, speakers who typically choose to listen to music during cognitively demanding tasks exhibit significantly lower F0 than those who do not. This interaction resembles the BinaryHoursXCondition interaction observed in EXP.2, where those who ChooseMusic (and likely listen to more music) are relatively more likely to show convergent behaviour to ambient pitch in speech production (Figure 4.8).

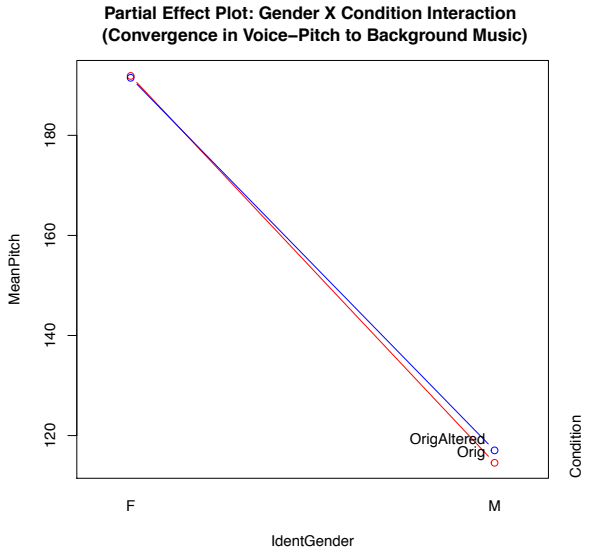

Figure 4.7 A partial effect plot visualizing the interaction between IdentifiedGender and Condition (Gender on the X-axis, Frequency in Hz on the Y-axis, and coloured lines indicating condition). While female speakers generally show little difference in mean F0 across conditions, male speakers appear to exhibit relatively higher voice-pitch in the lowered signal condition.

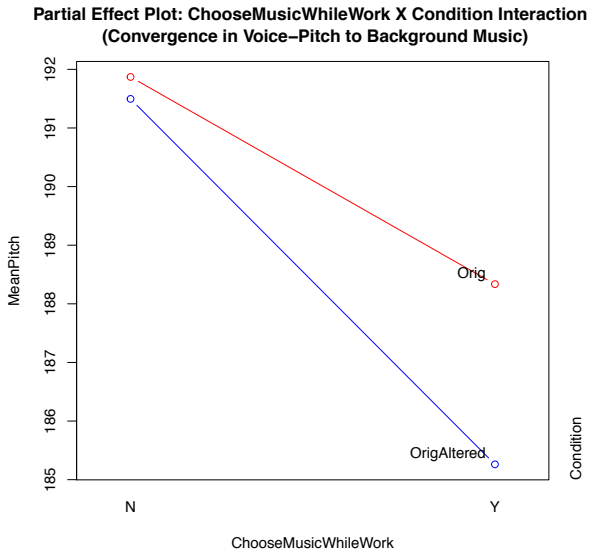

Figure 4.8 A partial effect plot visualizing the interaction between ChooseMusic and Condition (ChooseMusic on the X-axis, Frequency in Hz on the Y-axis, and coloured lines indicating condition). While speakers generally tend to exhibit lower F0 in the lowered pitch condition, the effect for speakers who choose to listen to music during cognitively demanding tasks reliably shows a relatively greater magnitude.

Before moving on to the speech-based analyses, it would make sense to first try to better understand these interactions. For example, in Figure 4.7 we see plainly that male speakers produce a lower F0 than female speakers in the lowered pitch condition; however, it is important to discern whether or not these gender-based groups are performing significantly differently *from themselves* across conditions. This can be tested in a straightforward way by once again subsetting the data, this time by Gender, and running comparable models to check for main effects of Condition. Subsetting the data in this way results in 1,512 observations from Male speakers, and 2,052 observations from female speakers. Fitting models to each subset that closely resemble the final music-based model, while removing any terms involving the IdentGender variable, shows that male speakers' productions exhibit significantly higher F0 in the lowered pitch condition (Est. 2.734, t = 2.669); this, however, is a partial effect and refers only to male speakers who do not choose to listen to music during cognitively demanding work. Releveling with ChooseMusicY as the reference indicates that male speakers who do choose to listen to music during such tasks do not perform differently across conditions (Est. -0.4121, t = -0.539). The female speakers, on the other hand, show no significant difference across conditions (Est. -0.5983, t = -0.640) for those who do not choose music; we do, however, observe a significant effect of convergence when we relevel female responses with ChooseMusic with "yes" as the reference (Est. -2.8273, t = -2.878). The ChooseMusicXCondition interactions for both Identified Gender categories have been plotted side-by-side as Figure 4.9, with female speakers on the left and males on the right. It appears that female speakers are generally predicted to converge, though show greater tendencies toward convergence when also choosing music. The male speakers, however, appear to show divergent tendencies when not choosing music, and no substantial effect when they choose to listen to music during cognitively demanding tasks.

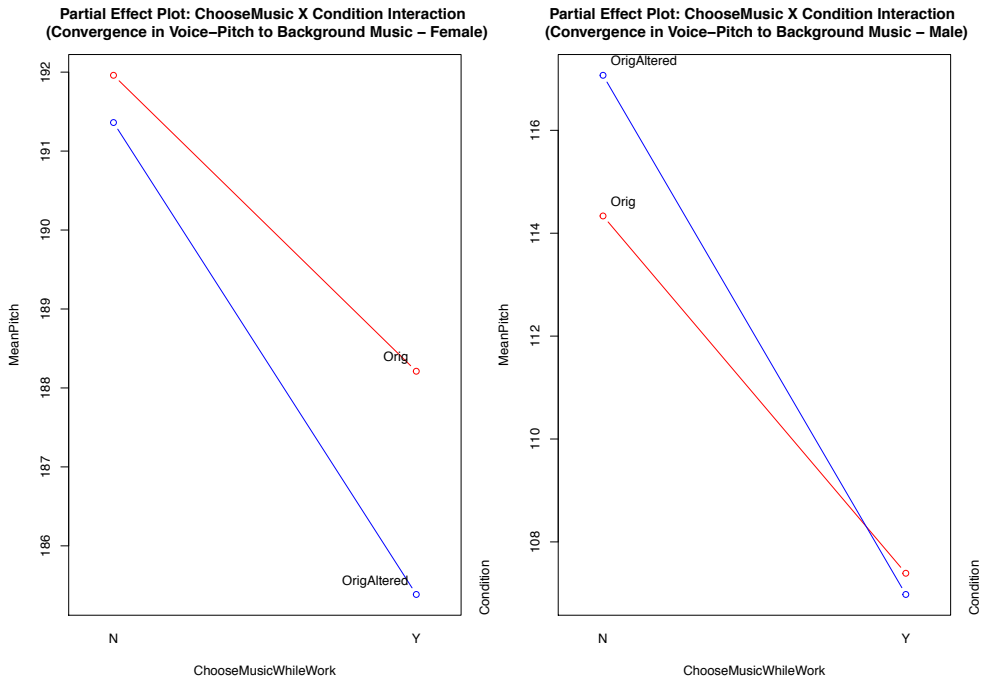

Figure 4.9 Two partial effect plots visualizing interactions between ChooseMusic and Condition with ChooseMusic on the X-axis, Frequency in Hz on the Y-axis, and coloured lines indicating condition). A plot for female speakers is available on the left, and one for male speakers on the right. While speakers generally tend to exhibit lower F0 in the lowered pitch condition, we do see divergent behaviour for male speakers who do not choose music during cognitively demanding tasks.

Similarly, we see that those who ChooseMusic and those who do not are performing significantly differently from each other in the lowered pitch condition. Once again, though, it is important to explore whether or not these groups are performing significantly differently from themselves across conditions. Subsetting the data based on ChooseMusic groupings results in 1,944 "yes" and 1,620 "no" observations. Fitting models to each subset that closely resemble the final music-based model, though here lacking any terms involving the ChooseMusic variable, shows that females who do not ChooseMusic do not perform significantly differently across conditions (Est. -0.4445; t = -0.514); releveling to test for males who do not choose music indicates a divergent effect for these speakers (Est. 2.880, t = 2.366). Conversely, female speakers who do ChooseMusic, appear to lower their voice-pitch in the direction of the manipulation (ConditionOrigAltered: Est. -2.840, t = -3.162); releveling to explore behaviour of the male speakers who choose music indicates no significant differences in voice-pitch across conditions (Est. -0.4842, t = -0.539). These interactions have been plotted below as Figure 4.10 with IdentGender on the X-axis, frequency in Hz on the Y-axis, and different coloured lines to indicate condition.

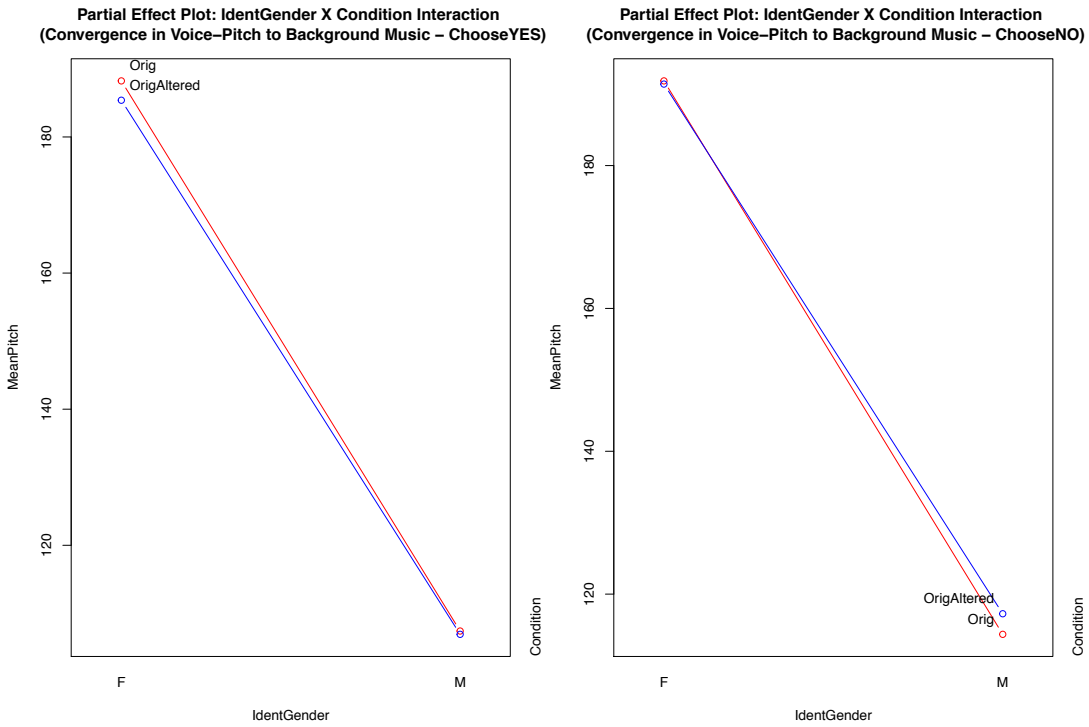

Figure 4.10 Two partial effect plots visualizing interactions between IdentGender and Condition with IdentGender on the X-axis, Frequency in Hz on the Y-axis, and coloured lines indicating condition). A plot for ChooseMusicYES is available on the left, and one for ChooseMusicNO on the right. The plot on the left indicates that when speakers choose music during cognitively demanding work, that females tend to converge while male speakers show no effect. Of the speaker who do not ChooseMusic, The females tend to show no effects, and the male speakers exhibit significant effects of divergence.

As a brief interim summary, when testing for entrainment to pitch-based variation in background music, the data from EXP.3 once again indicate that speakers are being influenced in complicated, but reliable ways to acoustically converge with or diverge from the signals they encounter. There appears to be a complex relationship involving Identgender and ChooseMusic when predicting these data, which may be compared to the BinaryHoursXCondition effect observed in EXP.2. That is to say, if we presume that speakers who ChooseMusic also tend to listen to more music per day, and that those who do not ChooseMusic listen to less music per day, then the predicted convergent/divergent effects observed for voice-pitch in background music seem reasonably consistent at this point (compare Figures 4.9 & 3.11). Simply, thus far it appears that variables implying greater music-listening habits are associated with convergence, whereas less music per day indicates divergence or at least reduced effect sizes.

4.7.2 Speech-based Mixed Effects Models

Next I describe mixed effects models fit to data collected during the speech-based treatments (3,563 observations). While testing for the possibility of convergence to ambient speech, a speakers' previous musical

training was included as suggested through the random forests. Again, IdentifiedGender and Condition were also included as control variables. All possible two-way interactions were explored. Random intercepts were included for Participant and Item, as were random slopes for Condition by Participant. The optimizer was again set to "bobyqa". The final model is provided below as Table 4.2.

	Estimate	Std. Error	t value
(Intercept)	189.652	5.652	33.553
ConditionSpeechAltered	2.425	1.566	1.548
IdentGenderM	**-75.268**	**8.674**	**-8.678**
ConditionSpeechAltered:IdentGenderM	**-4.891**	**2.405**	**-2.033**

Table 4.2 The final model predicting mean F0 in the Background Speech condition, using predictors from the random forest modeling. Significant effects have been bolded for convenience.

Once again, as expected, male speakers tend to speak with significantly lower voice-pitch than the female speakers. However, when testing for entrainment to ambient speech we also observe a significant interaction between Condition and IdentifiedGender (plotted below as Figure 4.11). It appears that male speakers exhibit relatively lower F0 in the lowered speech condition (i.e., convergence), whereas the female speakers appear to be diverging.

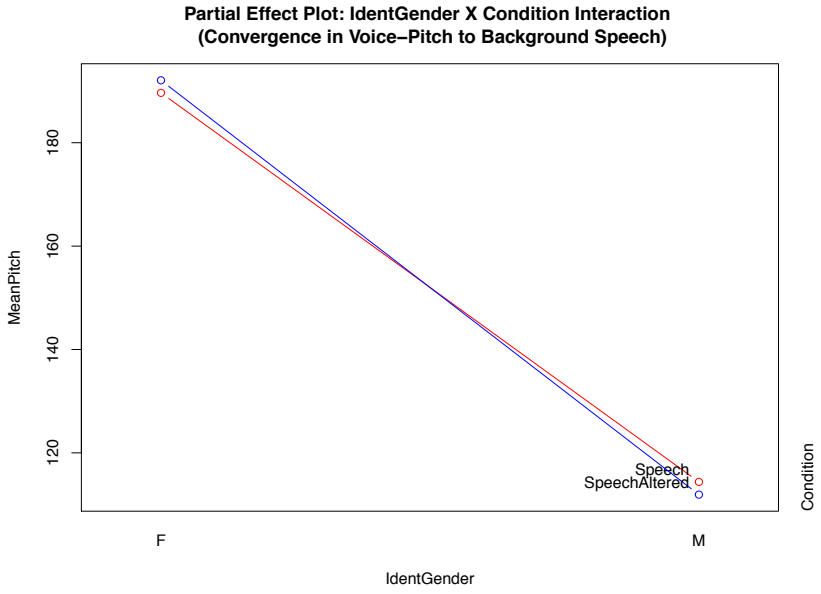

Figure 4.11 A partial effect plot visualizing the interaction between a speaker's identified gender and Condition (IdentGender on the X-axis, Frequency in Hz on the Y-axis, and different colours to indicate condition). While female speakers generally tend to invert the effect across conditions when exposed to background speech, male speakers are showing convergent behaviour.

This interaction may also be better understood by subsetting the data to explore whether or not the differences within groups are statistically significant. These speech-based data were therefore split by IdentGender to explore the IdentGenderXCondition interaction. Doing so left 2,052 observations from female speakers, and 1,511 from the male speakers. Fitting these data to a model much like the final speech-based model (though having removed all terms involving IdentGender) indicates that male speakers' productions exhibit significantly lower F0 in the lowered pitch condition (Est. -2.3308, t = -2.389). Testing similarly for productions with the female-based data, we see that the divergent trend is not statistically significant (Est. 2.372, t = 1.226).

Thus, in brief summary of the speech-based data analysis, it appears that speakers' productions can also be influenced by backgrounded speech, and that these effects seem rather different in drive than the music-based responses. Specifically, we find that a speaker's identified gender was the only viable predictor in this context, and that only the male speakers' responses to backgrounded speech reached significance. While female speakers also exhibited divergent behaviour, these changes in production did not reach significance. Importantly, with regard to methods, the only variable recommended through the random forests did not survive the introduction of random slopes in the mixed effects modelling. The only predictor observed to persist was IdentGender, which had been included as a control variable.

4.8 Discussion

This experiment has investigated phonetic convergence to pitch variation in ambient music, as well as the potential for phonetic convergence to pitch changes in multi-talker ambient speech. While EXPs 1 + 2 have already provided some evidence for convergence and divergence in voice-pitch to background music, this study also allows for similar, and directly comparable exploration of speaker-behaviour when talkers are forced to deal with noise from other speakers with whom they have no direct interaction. Moreover, this experiment explores how a speaker's susceptibility to such effects may be modulated and even predicted by both experiential and socially driven motivations. For convenience, a summary table of the effects observed within the study is available below as Table 4.3.

Condition (Manipulation)	**ChooseMusic**		**IdentGender**	
	Yes	*No*	*F*	*M*
Pitch - Music (was decreased)	Males: N Female: Y	Males: Y-inv. Female: N	n/a	n/a
Pitch - Speech (was decreased)	n/a	n/a	N	Y

Table 4.3 Summary of results for Experiment 3: "Y" indicates the presence of either a significant ChooseMusic or IdentGender interaction with Condition, whereas "N" indicates that no interaction was observed. The abbreviation "inv." is again used to identify situations where significant change was observed, though in a direction opposite to that predicted given the design of the test condition i.e., an inverted effect.

Once again we have seen that participants' speech production appears to be influenced in reliable ways by ambient noise, and that such influence can take place with both ambient music and backgrounded multi-talker speech. However, when the potential for systematicity in behaviours observed across conditions is considered, Figure 4.4 illustrates how speakers respond to the test conditions in every conceivable way. That is to say, some participants converged to both speech and music (n = 4); some diverged in both conditions (n = 7); exactly half of the remaining participants converged to background speech only and diverged from background music (n = 11); while the remaining half converged to background music only, and diverged from the backgrounded speech (n = 11). These data paint a clear picture of a fairly normal distribution, and alone may even suggest these behaviours were the product of chance. Though, the fact that both the random forests and mixed effects models recognized viable predictors for speaker behaviour in both conditions indicates that such behaviours were in fact far more principled.

Where in EXP.2 we saw that phonetic convergence to pitch in background music could be predicted to an extent through a speaker's performance in the Intensity condition (and vice versa), such correlations were not the case in the present experiment when attempting to make similar predictions using the background speech condition. In fact, the correlation tests between observed data and those simulated using random forest models from the opposite condition in this study indicate extremely poor predictive power. The above analyses from EXP.3 suggest that data gathered during either of the speech- or music-based treatments were not suggestive of performance in the other, even when further variables and forms of personal experience were taken into account through the random forest models.

Recall that EXP.2 involved a random forest analysis where t-values and correlations were recalculated after shifting observations to test for an inherent relationship between conditions (Section 3.10.1). Data suggested that much of the strong correlation was the product of a single control condition being used while generating t-values for both conditions; however, correlations were significantly stronger when observations were matched, showing that there was at least *some* degree of predictability gained through knowledge of speaker behaviours in other music-based treatments. In this study though, comparison of speech-based and music-based data suggest *no* correlation, which further suggests that convergence to ambient speech and convergence to ambient noise/music may be at least somewhat different processes.

I believe some differences in these forms of convergence may at least partially be rooted in attention. In the context of background speech, it is well-known that listeners are much more adept at filtering out noise created by many speakers than by a single speaker (e.g., Miller, 1947; Darwin, 2007; Cherry, 1953). If participants found a single voice particularly salient within the background signals (even if which voice was most salient varied from moment to moment), then we would expect relatively higher levels of distraction and perhaps a change in allocated cognitive resources for that speaker. Importantly, in the debriefing survey that followed each experimental session in this study, participants were asked if they had noticed anything about the sounds they heard through the headphones in the different conditions – this question was intentionally vague, and was designed to get at whether or not participants had consciously noticed acoustic manipulations. 73% of the participants who took part in this study either (1) Mentioned specifically that the speech-based conditions were more distracting than the background music,

or (2) Listed specific topics referenced by speakers within the speech-based stimuli. Despite compressing and scaling intensity for each of the four channels used to generate the speech-based background noise, as well as presenting two unrelated conversations simultaneously and at a remarkably low presentation volume of 45 dB(A), subjects were still processing these acoustic signals as well as the linguistic information encoded in those sounds (even if fragmented). Moreover, sufficient information was retained in memory after this processing took place that speakers were able to repeat it after the study had concluded (as long as 30 minutes after the fact). Clearly, background speech was an unexpectedly large draw on participant resources during this experimental task.

When considering the related cognitive processing differentiating these tasks (i.e., dealing with background speech vs. noise), Bronkhorst (2015) explains some of the *how*s and *why*s participants may have responded in this way, stating that most sounds "can be easily grouped—and subsequently selected—using primitive features such as spatial location and fundamental frequency. More complex processing is required when lexical, syntactic, or semantic information is used." Bronkhorst's statement suggests that different, or at least expanded cognitive work is required to process speech vs. other sounds, as different forms of semiotics are at play when processing speech as opposed to music notes, for example. In other words, while some listeners may ascribe abstracted meaning(s) to instrumental passages, these meanings are generally far less specific than the messages listeners take from spoken passages.

Thus, as mentioned above, it was not expected that musical consumption would influence convergence to background speech – and no such effects were observed. Moreover, while the random forest suggested musicianship would be important when predicting convergence to speech, this predictor did not withstand inclusion of the IdentGender variable in the mixed effects modelling (a variable which, notably, was not recognized through the random forest and had been included as a control). Taken together, these speech-based results suggest that background speech is likely processed differently than background noise in some important way(s), and further support the effects of music-related variables in the music-based conditions as real and not just random patterns in the data. As an avenue for future research, it would be valuable to find an analogue to some of the music-based predictors recognized as significant in a speech-based enquiry and test for comparable differences in the effects observed (e.g., how much time speakers spend in crowds daily).

Returning to the research questions driving this work, at this point relatively strong evidence has been collected across EXPs 1-3 that speech production can be influenced by ambient noise. Analyses across experiments have shown that speakers converge and diverge in multiple acoustic domains (typically converging with signals they are fond of, and diverging from signals they dislike or are relatively less fond of), and that both backgrounded music and speech can elicit such convergent and divergent effects. In light of this evidence, it seems extremely likely that the observed variation in speech is not just random noise, and is in fact structured and predictable.

However, when considering the second question, that is, whether or not there is meaningful variation across groups, and if any such variation might be predictable through social- and/or experience- based motivations, the answers are less immediately clear. While we find what seems to be structured variation within EXPs 1-3, the predictors recognized to explain the data differ from one study to the next. In other words, no clear drive for convergent/divergent behaviour has been identified through a single variable thus far. Recall, though, it was

suggested in EXP.2 that certain explanatory variables are thematically related, and perhaps may be involved in a dynamic relationship and/or are related to some unknown predictor which is the primary drive for such effects. In light of the current analysis I will argue that these options need not be mutually exclusive, and that perhaps *the relationship itself* connecting these variables is in fact partially driving some of the behavioural changes observed throughout this work.

Most of the social- and experience-based information collected for use as predictors in this dissertation maintain some grounding in musical experience/consumption, so it should come as no surprise that these variables pattern together in interesting ways. While this thematic interrelatedness among predictors can be taken as a potential limitation of the these experiments, it may also be recognized as a strength; varied results would be more concerning if, for example, "Handedness" came up as the best predictor in one experiment and then "Block" or "PresOrder" in the next. The fact that the most robust predictor in EXP.2 was the BinaryHours measure (which harmonized well enough with results from EXP.1 after looking at the distributions of Musicianship and their listening habits) and ChooseMusic for the music-based condition in EXP.3 suggests a relatively clear pattern. For a moment disregarding the gender-based effect, we find that female speakers in EXP.3 who ChooseMusic while working through cognitively demanding tasks exhibit significant convergence-based effects in the music condition. It seems very likely that these speakers would also listen to more music generally than speakers who do not choose music in these contexts. Conversely, we find that the male speakers in EXP.3 who do not ChooseMusic while working were diverging from the lowered pitch manipulation. To test this relationship, I have plotted the distributions of BinaryHours against ChooseMusic for participants in EXPs 1-3, which is available below as Figure 4.12 (recall that the cutoff in EXP.1 was two hours instead of one). The plot shows clearly that speakers who listen to relatively more music are also more likely to ChooseMusic, whereas speakers who listen to relatively less music are more likely to *not* ChooseMusic.

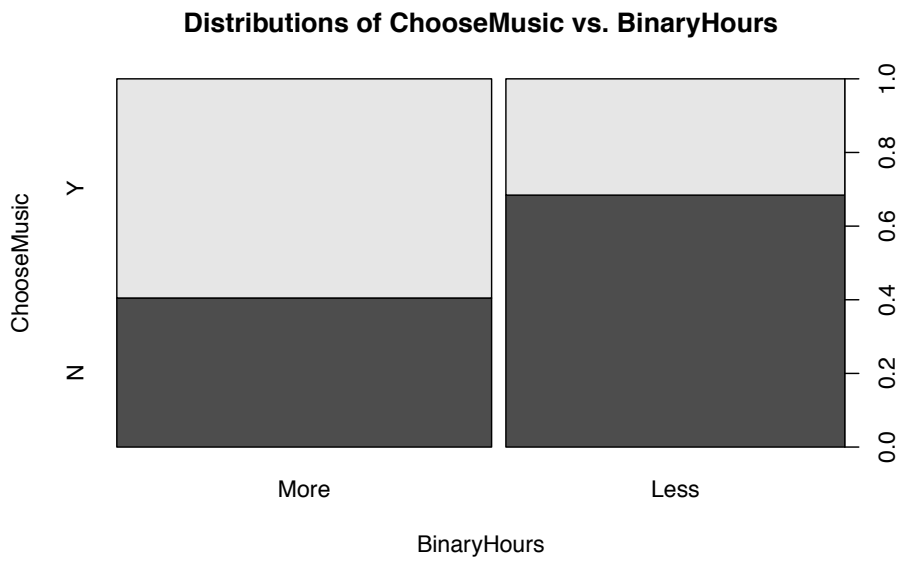

Figure 4.12 Distributions of ChooseMusic compared to those of BinaryHours across EXPs 1-3. As suggested, we see that speakers who choose music during cognitively demanding tasks tend to listen to relatively more music, and that those who do not choose background music in this context listen to less music generally.

Thus, the descriptions of convergent and divergent effects observed in EXP.3 seem much like those recognized in EXPs 1 + 2, where speakers who listen to relatively little music (as measured through the BinaryHours variable) were also found generally to exhibit divergent behaviour and speakers who listen to more music are likely to converge. These effects remain consistent with social motivations and alliances as a drive for convergent and divergent behaviour. Much like with communicative speech, if talkers hold positive opinions of whatever might be influencing them they are more likely to converge. On the other hand, talkers who hold less positive opinions are more likely to show no effect or to diverge, much like in the effects described above.

Next, returning to the interaction between IdentGender and ChooseMusic, we see patterns in the music-based data that also look very much like effects observed previously in this work. Specifically, in Experiment 2 we saw that male speakers were predicted to diverge from the lowered pitch manipulation, whereas female speakers exhibited convergent tendencies that did not reach significance. The analysis in EXP.3 also found male-based divergence and female-based convergence, but further suggests that this effect driven by the speaker's listening habits is at times mediated by a speaker's gender. It is unclear exactly why the gender split is influencing speakers in this particular way, but gender-based performance does appear relatively consistent across studies.

Thus, musical consumption and gender both appear to influence convergent and divergent behaviour, and it seems to be the case that these variables can influence speaker-behaviours together and in isolation. Indeed, much like was argued in EXP.2, it seems both social- and experiential motivations are consistent with the altered speech productions observed in Experiment 3 (as well as the previous studies in this collection). Because convergence effects are known to be somewhat irregular (cf. Pardo, 2013), perhaps the specifics of which variables aid in

predicting some of these effects may be of less consequence than the themes that relate them. Indeed, of equal importance seem the social factors and experiences that appear to modulate them.

Focussing specifically on the speech-based analyses for a moment, it is remarkable that the random forests recognized any of the thematically music-focussed variables as important when explaining variation in this condition – a condition that we have seen appears to differ from convergence to background music in some significant ways. Two points of note come to mind immediately: The fact that a significant entrainment effect has now been observed for the speech-based stimuli when presented at ~45dB is important because the known threshold for Lombard responses to speech-based noise has been recognized previously as ~55 dB(A) (Lazarus, 1986). The reason for differences in convergent/divergent behaviour by gender is unclear, though such gender-based effects are encountered in this line of research often enough (e.g., Drager, Hay & Walker, 2010). However, *any* entrainment-based effects at this presentation level add further support for the entrainment-based processes observed in this dissertation as distinct from Lombard speech. When considered along side the entrainment to intensity observed in EXP.2, also below the known Lombard threshold, it appears that the experiments comprising this dissertation serve as the first steps toward understanding an exciting new area in speech production research which certainly warrants more work.

A second point involves the random forests identifying Musicianship as a viable predictor for effects related to the speech-based treatments – recall, though, that this predictor did not survive the mixed effects modeling process. In fact, the only predictor to remain significant in this model was IdentGender, which had primarily been included as a control variable. In conjunction with some of the IdentGender-based findings from EXP.2 (where this variable was also not recognized through the random forests, but was found to be highly significant in the mixed models), it appears that random forests may be most useful in an exploratory sense, where other variables motivated for inclusion for various reasons are also worth considering in complementary modeling. That is to say, while random forests have been helpful throughout the analyses described in EXPs 2 + 3, other tests were required to flesh out a better understanding of the data. Indeed, other variables motivated through previous research must not be omitted from analyses solely because they were not recognized as important by the random forests. Use of multiple statistical tests in tandem is therefore well supported at this point as a sensible way to proceed through necessarily murky waters.

Finally, in light of the seemingly different processes involved in convergence to ambient speech vs. ambient noise, it seems worthwhile comparing the magnitude of these effects. Because forms of phonetic convergence observed during communicative speech are at times consciously recognizable by listeners (e.g., Pardo et al., 2012), it was unknown whether or not the magnitude of convergence to background speech may also be more extensive than phonetic convergence to background music. Indeed, convergence to- and divergence from pitch manipulations in backgrounded music throughout the present work have been consistent, though estimates have been modest across analyses.

Despite the different draws on cognitive resources which appear to be involved in these two forms of convergence, the magnitude of the effects observed appears to be largely similar across conditions (mostly within the single-digit Hz range). These similar realizations of convergence/divergence (despite some apparent differences in

processing) seem potentially important to theories of speech perception; specifically, such similarity may be taken to suggest the processes involved may not be wholly different. Considering the claims of Bronkhorst (2015) regarding differences in information-processing with regard to sound vs. meaning, it is possible that such convergence takes place prior to any linguistic processing rooted in semantics, which would likely result in somewhat similar changes to speech through various forms of background noise. We have seen throughout EXPs 1-3 that convergence and divergence from background music appears to be mediated by personal experience at some point in processing – data from EXP.3 suggest that speech-based convergence/divergence is also further shaped by experience, though by *different* experiential drives.

4.9 Conclusion

This experiment has explored the notion of phonetic convergence to both background music and to backgrounded multi-talker speech. I have developed and discussed methods to explore these avenues in ways that are directly comparable in order to better understand the cognitive processes that may influence this type of convergence in speech production, as well as the magnitude of effects across these (at least somewhat) different situations. The data analyzed within this study indicate that convergence to speech may be a different process than convergence to background music, at least in certain respects; and, importantly, effects recognized in the speech-based treatments in this study have played a key role in distinguishing entrainment effects from other speech-in-noise processes, such as the Lombard effect. Most importantly, the present study has collected further evidence to support speakers' productions as influenced by ambient sound in statistically significant ways. The above analyses support both social and experiential-based motivations for effects of convergence and divergence in speech, where once again we see that reduced musical listening habits are predictive of divergence and increased listening habits predict convergence – though, both effects are further influenced by speaker gender. While much work must be done to understand the specifics of how human speech is influenced by ambient sound, speech-based or otherwise, this enquiry has shown that follow-up work on acoustic convergence and divergence to ambient noise is warranted.

CHAPTER 5: A Unified Analysis

5.1 Introduction, and Summary of Previous Findings

Together, the three experiments described above provide relatively strong evidence for human speakers converging and diverging acoustically with certain aspects of ambient noise during speech production. Results across studies, however, were not immediately easy to interpret together; variables recognized as important predictors differed across studies, but were later recognized to share a thematic relation. For ease of comparison, summary tables illustrating the effects observed from each study can be found below as tables 5.1-5.3. Recall that in all tables "Y" indicates that speech was generally altered to exhibit characteristics of the acoustic manipulation, and "inv." describes situations in which speech was generally altered in the direction opposite to that of the manipulation. "N" simply describes situations where no significant changes were observed. Relevant effects have been bolded in cases where multiple comparisons took place.

Acoustic Dimension	**A/B Comparison (no predictions)**		**B/C Comparison**		**PropChange**	
	Mus	*NoMus*	***Mus***	***NoMus***	*Mus*	*NoMus*
Pitch (was decreased)	Y - decrease	Y - decrease	**Y-inv.**	**N**	Y-inv.	N
Intensity (was increased)	Y - increase	Y - increase	**Y**	**Y**	Y	N
Tempo (was decreased)	N	Y - decrease	**N**	**N**	N	N

Table 5.1 Summary table describing results from EXP.1. The B/C comparison (unaltered music in "B" vs. the lowered-pitch/intensity conditions in "C") has been bolded as the analysis most directly comparable with those of EXPs 2+3.

Acoustic Dimension	HoursMusicPerDay	
	1hour or less	> 1 hour
Pitch (was decreased)	**Y-inv**	**Y**
Intensity (was decreased)	**Musicians: N** **SMs: Y-inv** **Non-musicians: Y**	**Musicians: Y** **SMs: Y** **Non-musicians: Y**

Table 5.2 Summary table describing relevant effects observed in EXP.2.

Condition	ChooseMusic		IdentGender	
	No	*Yes*	*F*	*M*
Pitch - Music (was decreased)	**Males: Y-inv.** **Female: N**	**Males: N** **Female: Y**	n/a	n/a
Pitch - Speech (was decreased)	n/a	n/a	**N**	**Y**

Table 5.3 Summary table describing results from EXP.3 for both music- and speech-based conditions. Note that different effects and predictors were recognized in these analyses, and that effects in the music-based condition were driven by variables related to those observed in EXPs 1 + 2.

As mentioned in EXP.3's discussion (Section 4.8), the primary effects observed across studies do appear related through participant listening habits (recall that the distributions of Musicians and HoursMusic are very similar across studies; see Figure 4.12) – though, these trends have all been recognized through isolated analyses. While unlikely at this point, it is still possible that the observed effects were recognized through Type 1 errors driven by insufficient or flawed data, or methodological differences across studies. To ensure the legitimacy of these effects and relationships across experiments, it seems prudent as a final step in this investigation to explore data collected during the three studies in a single, unified analysis. Specifically, these efforts will test whether or not the observed effects withstand a more robust data set; it would be predicted that patterns recognized in previous analyses would disappear if they are, in fact, driven by noise. This chapter is dedicated to such a unified analysis, where the statistical tests described below further support the effects observed throughout this work as genuine (that is, relatively less time spent listening to music daily is predictive of divergence, whereas relatively more time spent listening to music daily is predictive of convergence).

Approaching a unified analysis proves difficult in this situation however, as the nature of the studies and data collected throughout this work are less consistent than would be ideal. Most notably, data collected during EXP.1 are comprised of time-series observations associated with acoustic manipulations over time within a single

(relatively longer) trial. Data collected during EXPs 2+3 however, involve many independent observations extracted as mean values from each of many individual trials, each associated with global manipulations that were grouped and experienced by Block/Condition.

While the studies comprising the present work are fundamentally different in some important ways, the resulting data have been transformed so as to make observations relatively more comparable. Pitch is the only acoustic dimension tested in all three studies – as a result, this unified analysis focuses exclusively on convergence to ambient pitch in speech production. Much like the analyses described in EXPs 2+3, this unified analysis begins largely in the realm of t-values by Condition for each participant. These values have been calculated already, for the most part, during the analyses outlined in EXPs 2+3. Comparable values need only be generated for data collected during EXP.1 to prepare a unified dataset. Recall that each condition in Experiment 1 was divided into *Sections* (see Figure 5.1 below):

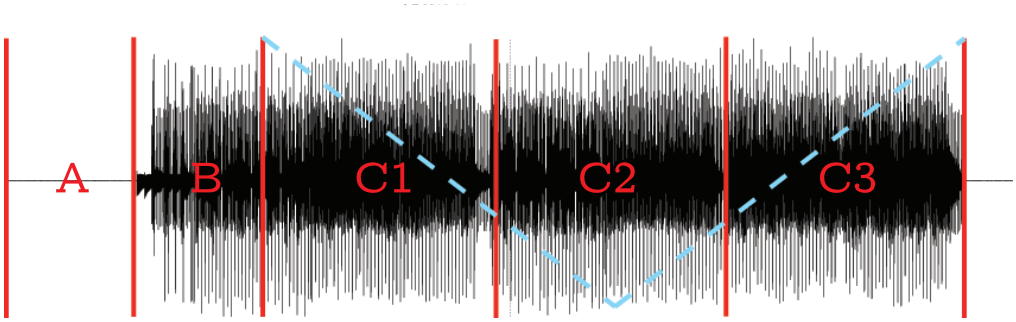

Figure 5.1 Within each each Block (or Condition) the speaker's productions were split into 5 Sections representing the degree of manipulation in the stimulus at a given point. Section A denotes speech produced before the introduction of background music, whereas Section B shows the introduction of background music before acoustic manipulation begins. Sections C1, C2, and C3 describe the onset of manipulation, the manipulation reaching its target/maxima, and the return to onset values, respectively.

The t-values used in EXPs 2+3 were generated by comparing observations from a baseline condition to those from a globally lowered-pitch condition. While not perfectly comparable, a similar value is generated for each participant in EXP.1 by extracting and then comparing observations from Section B (unaltered music) and Section C2 (where the lowered pitch manipulation reaches its maxima) in each participant's pitch condition. To this end, unpaired t-tests were run for samples of unequal size to compare observations from these sections for each speaker – much like in the previous analyses – in order to generate a normalized measure of effect size and direction that might be used to compare performance in EXP.1 to the later studies (note that unpaired t-tests were most appropriate in this context because there were unequal numbers of observations in Sections B and C2, and the condition was comprised of a single article as reading material as opposed to multiple comparably matched trials). All t-values were centered and scaled by study prior to combining them in order to avoid skews by data set.

5.2 Analysis

With t-values now available for each participant's pitch condition(s) in all studies, we can first explore the possibility of convergence in voice-pitch through random forests. Please note that the analyses described in EXP.3 support convergence to background speech as somewhat different than convergence to background music, so the analysis below excludes t-values from EXP.3's speech-based condition (80 observations in total are explored here). Once again, a forest was grown using Participant[19], IdentGender, Age, Musical training, ChooseMusic, and HoursMusicPerDay to predict t-values. Experiment was also included as a control variable (as a three-level factor). Mtry was set to 2 and ntree set to 5000, resulting in the following variable importance plot (figure 5.2):

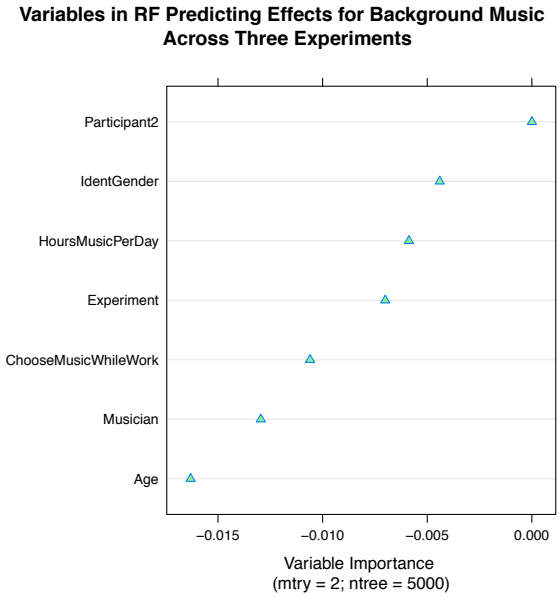

Figure 5.2 A variable importance plot resulting from the random forest fit to t-value-transformed data from EXP.1-3. When considering data from all three studies, no variables are recognized as important; though, this lack of effects seems more likely the product of how data for EXP.1 were collected (and subsequently transformed).

Looking at Figure 5.2, with all values either at or below the '0' mark, it is quickly apparent that none of the variables supplied appear useful when predicting a speaker's pitch-based variability across the three studies. This was not completely unexpected however, as the data representing EXP.1 are not directly comparable to those gathered during EXPs. 2+3, nor are the methods used to transform those data. To ensure that those data were not contributing anything meaningful to this particular analysis, a forest was fit to EXP.1 t-values only (Figure 5.3(left)) and no variables were recognized as viable predictors of the data; with only 15 observations and all variables with variable importance of '0', it is assumed there were insufficient data in this context to extrapolate any coherent trends. As a

[19] In this context Participant is a hybrid variable incorporating both a speaker's identification number and the experiment ID as a single predictor. This was done to differentiate observations from participants across studies (e.g., participant 1 in EXP2 from participant 1 in EXP3).

result, it was decided that these observations should most likely be dismissed due to the problematic data from which they were created (i.e., autocorrelation) and the noise they add following the t-value transformation. These observations were therefore removed from the dataset, leaving 65 observations in total. Another forest was generated under otherwise identical conditions, fit to the remaining observations from EXPs 2+3, and this restricted unified forest resulted in the VIP available below as Figure 5.3 (right):

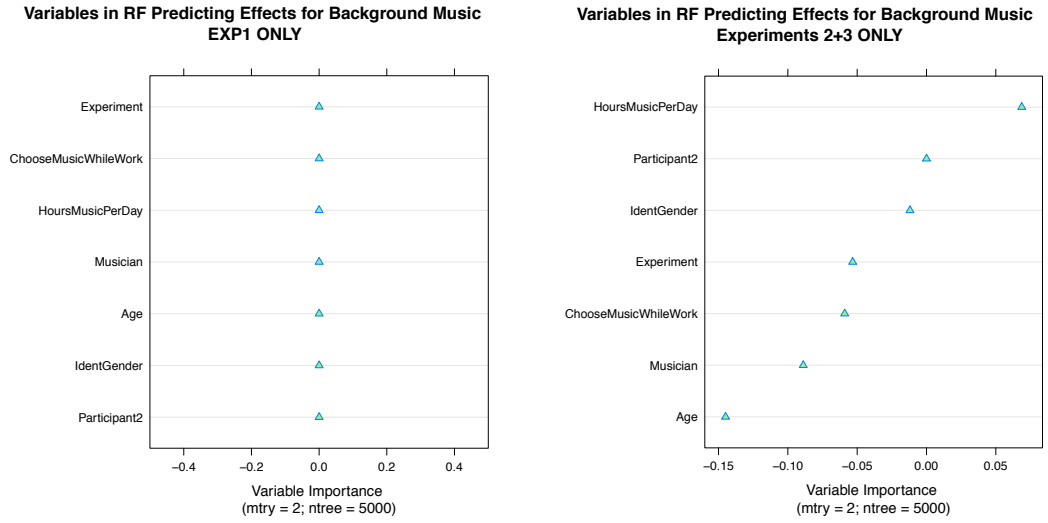

Figure 5.3 (Left) A variable importance plot generated from the random forest predicting t-value-transformed data collected during EXP. 1 only. (Right) A variable importance plot generated from the random forest predicting t-val-transformed data collected during EXP. 2+3.

Here, we see that the self-reported HoursMusicPerDay variable (included as BinaryHours) is again showing predictive power when accounting for variation in voice-pitch across the two studies. As in EXPs 2+3, results from the Random Forests were next used to feed a linear regression. Though, in this case models exclude random effects and were fit to the scaled, t-value-transformed data instead of raw data collected during music-based conditions in EXPs 2+3 (n=65). HoursMusicPerDay, Experiment and Identgender were explored as predictors, where IdentGender was also included here due to its importance shown in previous chapters. Models were fit using a backward stepwise selection procedure where subsequent models were compared using ANOVA tests in R. All potential two-way interactions between HoursMusicPerDay, Experiment, and IdentGender were also tested, and the final model is available below as Table 5.4.

	Estimate	Std.Error	t-value	Pr(>|t|)	
(Intercept)	-0.3401	0.1836	-1.853	0.06874	.
BinaryHoursOneOrLess	**0.9791**	**0.3103**	**3.156**	**0.00249**	**
IdentGenderM	0.5591	0.3483	1.605	0.11362	
BinaryHoursOneOrLess:IdentGenderM	**-1.3509**	**0.4922**	**-2.744**	**0.00795**	**

Table 5.4 The final model summary describing a linear regression fit to scaled t-values based on data collected during music-based conditions in EXPs 2+3. Significance codes are available in the right margin, and significant effects have also been bolded for convenience.

This model confirms the predictive power of HoursMusicPerDay through a main effect of BinaryHours. It appears that, across the two studies, the general trend for female speakers who report listening to more than 1 hour of music per day is to converge with pitch manipulations in background music, where those who report listening to one hour of music or less per day tend to exhibit divergent behaviour. Note, though, that the model summary is describing a partial effect, and that this summary does not describe the behaviour of speakers identifying as male. Subsetting the data by IdentGender leaves 40 observations for the female speakers, and 25 for the male speakers. Fitting similar linear regression models while removing all terms and interactions related to IdentGender indicates that this main effect hold for the female speakers (Est. 0.9791, t = 2.847) while male speakers show no reliable effect of the treatment (Est. -0.3718, t = 0.230). It appears that the trends encountered in the previous experiments are holding somewhat through the unified dataset, at least insofar as we encounter gender-based differences and the female participants, who make up the majority of these data, are performing as would be predicted based on the findings described in the individual analyses.

These effects can be better understood as the observed significant interaction between BinaryHours and IdentGender, which supports the possibility of a dynamic relationship between multiple variables in shaping these convergent and divergent behaviours. The interaction has been plotted below as Figure 5.4, again with BinaryHours on the X-axis and scaled t-values on the Y-axis – though, here the addition of coloured bars indicates Identified Gender. As described above through the subset analysis, we see through this figure that male speakers who listen to more than one hour of music per day trend toward converge in voice pitch, and those who report listening to more than one hour of music per day trend toward divergence – though, these differences are non-significant. Conversely, female speakers exhibit significant effects of convergence for those who listen to relatively more music per day, where female speakers who listen to relatively little music per day exhibit significant effects of divergence.

It should be noted however that due to the availability of participants these cells are not balanced, and there are nearly twice as many female participants represented in the data than there are male speakers. Gender-based effects might therefore be considered with caution. There is no clear explanation for why speaker-gender may negate the influence of HoursMusic in this way, though such an interaction does support the potential for a dynamic

relationship between gender, listening habits, and perhaps even musical experience (cf. EXP.1) when modeling convergence and divergence to background noise in speech production.

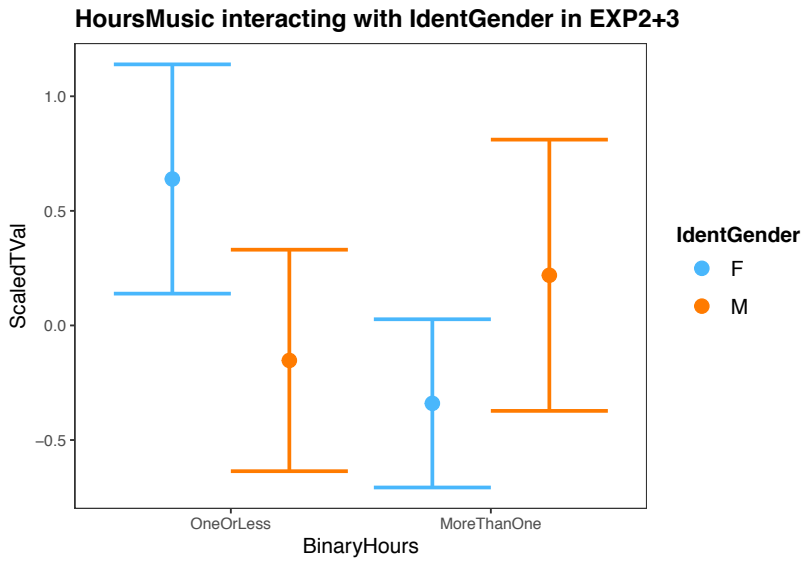

Figure 5.4 Plotting the interaction between (binary) Hours Music Per Day with Identified Gender observed when fitting a linear regression to scaled t-value-transformed data collected during EXPs 2+3. This figure supports a complex relationship between these two predictors, where music-per-day has opposite effects in light of gender differences. We see that females who listen to more than one hour of music per day an also the males who listen to one hour a day or less are associated with negative t-values (i.e., convergence). We also find that females who listen to one or less hours of music per day and the males who report listening to more than one hour per day of music are generally associated with positive t-values (i.e. divergence).

5.3 Discussion

In light of the preceding analyses, unified and otherwise, it appears that results have come together to tell a relatively coherent story. Where we were previously unable to identify a specific variable to predict convergence and divergence to ambient noise across all three studies, the above unified analysis indicates that the BinaryHours and IdentGender variables both maintain significant predictive power when models are fit to data from EXPs 2+3. Indeed, we have also seen that comparing distributions of BinaryHours and those of Musicianship yields similar results as well (Figure 3.17), thus largely harmonizing findings across EXPs 1-3.

Exploring these findings, then, in light of the research questions posed at the beginning of this work, there is strong evidence to support that the productions of (at least some) human speakers are indeed influenced by ambient noise in reliable ways. While speakers do not always react similarly to different acoustic treatments (i.e., always converge *or* diverge), the patterns of convergence and divergence observed across studies indicate systematicity in these events; indeed, the above unified analysis further supports these findings as principled and *not* the product of Type 1 errors. In other words, because certain effects and certain variables have been identified repeatedly

throughout the present work as statistically significant, and were again supported through an expanded data set in the unified analysis, these effects appear to be real and not just random variation mistakenly recognized in the data.

With such support in mind, we can now revisit the second research question regarding meaningful variation in production, and potential predictability of that variation through social and experiential motivation. Having provided reasonable evidence that speakers do, in fact, at times converge with and diverge from ambient sounds in speech production, this work has identified multiple experiential variables capable of predicting speaker behaviours across studies to some extent. That is to say, while trends were recognized in each individual analysis, the importance of speaker listening habits has been supported continually through significant effects of BinaryHours and/or ChooseMusic, and the repeated significance of IdentGender shows that a speaker's identified gender also has an important role to play in shaping speech production in this context. This role is not particularly clear at present, though gender-based differences are not unfamiliar in the accommodation literature (for example, Hay & Drager (2010) also describe contexts where female speakers show effects of convergence where male speakers showed no effect). Crucially though, while gender-based effects are often the product of physiological differences, I can think of no known physiological difference(s) that could account for the related variability observed throughout this work. I therefore categorize these gender-based effects also as forms of experiential motivation, as such conditioning would more likely be attributable to contrasting experiences while living and learning while identifying as male vs. identifying as female.

Finally, the unified analysis provides continued support for the drive of such convergent/divergent behaviours as at least partially driven by social-effects – that is, an expression of alliance and/or opinions and attitudes. Much like forms of convergence observed in previous sociolinguistic studies, where speakers are found to converge more with conversation partners they enjoy or hold positive opinions of, and diverge from speakers they dislike or are less inclined to align with for one reason or another, similar patterns have been recognized in convergence and divergence to backgrounded music. We have seen that talkers who listen to relatively less music often diverge acoustically from the background music, whereas speakers who self report as listening to relatively more music per day on average most often exhibit convergent tendencies in their speech production. It appears that speakers' fondness for music is reflected in the form of entrainment realized in their speech production.

This finding raises an interesting question about the generalizability of these effects, and how talkers may respond to other forms of ambient noise. Early in this work similarities were outlined relating acoustic components of speech and music, where the primary building blocks of message transmission in both of these contexts were for the most part identical. Indeed, as shown above, a large portion of the speakers sampled within this work have regular interaction with music, at least to some degree, on a daily basis. Future works might explore whether or not convergence and divergence to forms of background noise involving less or no semiotic mappings may differ in their realization. It would be very interesting to explore whether all speakers diverge from the sound of, for example, a jackhammer on the street which, as a generally annoying sound would likely be predicted based on the work described above – or, if perhaps other variables reflecting speaker attitudes might be brought to light to better predict/recognize convergence and divergence in that specific context. In other words, are the behaviours I have associated with music-listening habits indicative of differently developed cognitive mechanisms involved in

listening, or might instead musical-listening habits be associated only with performance in the presence of background music? Differences in responses across speech- and music-based conditions might suggest the latter, though this interpretation is by no means conclusive.

Thus, in brief summary, at this point I have provided support for an influence of ambient noise on speech production, where effects appear to be manifest as forms of convergence and divergence. These behaviours are often statistically significant within-condition for a speaker, though participant responses (i.e., convergence or divergence) are not always uniform across different treatments. Specific experience-based predictors have been identified through multiple analyses, which together implicate a speaker's listening-habits as playing an important role in shaping the realization of entrainment to background music. With regard to theory and the drive/benefit of these changes to speech production, we have also recognized that the observed forms of convergence and divergence strongly resemble behaviours recognized in previous sociolinguistic efforts which reflect speaker attitudes and alliances (e.g., Drager, Hay & Walker, 2010), where certain effects even more directly parallel some of the less interpretable gender-based effects observed within the present work (Hay & Drager, 2010). Therefore, the present work supports both social and experiential motivations for convergence and divergence in human speech. The chapter that follows serves as a general discussion, first summarizing the primary benefits gained through the individual studies described in this dissertation, before finally addressing the research questions driving this work in more detail.

CHAPTER 6: General Discussion and Conclusions

6.1 The Beginning of the End

In the preceding chapters I described a series of three experiments designed to explore various forms of acoustic convergence to ambient noise during speech production. Each of these studies was designed as a speech-in-noise reading task where, specifically, I tested for acoustic-phonetic entrainment to background signals in three dimensions (Pitch, Intensity, Speech rate) while talkers experienced background music and/or background speech – the latter experienced as multi-talker babble. The experiments themselves were developed in order to explore two specific research questions motivated by previous works (for such motivations please see Chapter 1):

1. **Are speakers reliably influenced by acoustic characteristics of ambient noise, such that acoustic characteristics of their speech become systematically more or less like corresponding characteristics of that noise?**

2. **Is there meaningful variation between individuals or groups in terms of convergent and divergent behaviour – and, if so, to what extent might this variation be predictable through social- or experience-based variables?**

I will therefore next briefly recap the individual experiments, where each will be discussed in light of these research questions. Following these summaries, I will then discuss the implications of the present work.

6.2 A Brief Summary of Experiments and Findings

As mentioned above, the present work involved a series of three experiments. These studies were designed specifically to investigate the research questions outlined above, and therefore evolved to do so more effectively over the course of this project. Each experiment utilized a speech-in-noise reading paradigm, where talkers read over background music composed specifically for this work and/or multi-talker speech-based signals. In EXP.1 talkers produced connected passages selected for similar style, complexity, and content, which had been excised from nature magazines; in EXPs 2 + 3 speakers reproduced a list of words and phrases comprised of only sonorant phones in order to avoid analytical issues recognized during EXP.1.

The scope of this work was narrowed progressively with subsequent studies. Specifically, EXP.1 was designed to explore potential convergence in Pitch, Intensity, and Speech rate to background music; EXP.2 was largely a replication of Experiment 1, though tested only for convergence in pitch and intensity to background music (though, with a slightly altered methodology); and EXP.3 adopted the altered methods of EXP.2 to explore

convergence in voice-pitch only, though introduced speech-based treatments in addition to the background music used earlier. Through this increasingly focused enquiry, I have gathered evidence supporting an influence of background noise on human speech production. The specifics of results varied by study and initially were somewhat difficult to understand when considered together. Simply, no single predictor had been identified that could reliably explain variation in speech across all data collected during the present investigation. Patterns and themes however, emerged soon after and these apparently disparate effects were largely harmonized through effects tied to speakers' listening habits. I will next briefly summarize the most coherent results across studies before moving on to discuss some implications of the present work.

6.2.1 Experiment 1

Experiment 1 served as the first known attempt to investigate acoustic convergence to backgrounded non-speech noise as a possibility. With no directly related previous work to guide this enquiry, it seemed reasonable to impose manipulations gradually over time to test whether or not speakers would entrain to acoustic envelopes progressively (i.e., changing in real time). However, analytical issues stemming jointly from the time-series data and low participant numbers made analysis difficult and less straightforward than was hoped for. Sensible statistical methods had been selected, but the above-mentioned issues resulted in a low-power analysis – important findings still did come from this study however. It was entirely possible that the background signals could have had no effect on talkers' speech production; if this were the case, there would have been little support to continue with this line of research. The fact that statistically significant convergent/divergent behaviours were observed in two of the three acoustic dimensions tested (and, admittedly, it is possible that the analytical methods selected were not sensitive to Tempo-related convergence) at the very least warranted a replication study. Finally, having recognized that the background music did appear to reliably influence speakers to alter production patterns, a precedent was set in this new line of research insofar as a relatively low presentation volume of ~45 dB was sufficient to elicit these changes in speech production. Thus, while not perfect, the methods developed and used in this study were also not completely ineffective ways to test for entrainment to ambient noise, though methods would be improved in specific ways for the studies that follow.

With regard to the first primary research question, this experiment has provided some evidence that speakers are reliably influenced by background music, and that their productions do appear to systematically become more (and sometimes less) like that noise acoustically at times. It was not initially clear exactly why some participants were observed to diverge during the pitch-based treatment, while all participants were found to converge during the intensity-based treatment. It seems likely that the low participant numbers played a detrimental role in this analysis, where effects were likely not wholly recognized in this analysis (cf. EXPs 2+3). Simply, we only saw part of the picture because the sample size was just too small. Moreover, in addition to the less than ideal sample size, altering methods in the studies which follow EXP.1 to test treatments by Block (instead of gradually within a trial) would eventually provide a more robust analysis due to increasing the number of independent observations per speaker, therefore aiding the more accurate recognition of speaker trends.

While analytical methods most likely played some role in the disparate effects observed across conditions, it seems also possible that all speakers converging to raised intensity was a product of the Lombard effect. Recall that Lazarus (1986) describes the known threshold for eliciting Lombard speech as ~45 dB for non-speech noise; it seems possible that all speakers "converged" with the raised and then lowered intensity of this treatment because they were exhibiting signs of Lombard speech. Importantly, this trend can be compared with intensity-based effects observed in EXP.2, where speakers were found to both converge and diverge with a manipulation where intensity was lowered below this estimated threshold. While certain aspects of these processes are similar in their realization (that is, convergence vs. Lombard), data from EXPs 1-3 indicate they are most likely distinct. Perhaps the reason that all speakers converged to intensity in EXP.1 is because the Lombard effect overpowers convergence-based effects in this context (cf. EXP.2). Such unequal influence on speaker behaviour would make sense, as the Lombard effect is most often argued to exist as a compensation in speech which is driven by increased intelligibility. I have argued throughout this work that convergent and divergent behaviours to ambient noise closely resemble convergence/divergence observed in previous communicative speech-based studies, where in this context such behaviour is typically associated with an expression of speaker attitudes and alliances. The fact that a speaker would unconsciously prioritize intelligibility and effective communication over externalized attitudes – where the latter may not be expressed effectively if communication channels are in question – makes perfect sense. This behaviour can be contrasted with later performance in EXP.2, where speakers exhibit both convergence and divergence to intensity-based manipulations; if the speakers are no longer susceptible to the Lombard effect due to relatively higher presentation levels, and effective communication was therefore no longer in question because the presentation level of background noise was below the known threshold, then speakers would have more freedom to express secondary information (e.g., attitudes and opinions) through all available channels because the primary information (i.e., the semantic information encoded through lexical items and syntax) is likely to be understood.

Thus, when considering the second research question regarding meaningful variation and predictability via social- and experienced-based motivations, data from Experiment 1 appear to implicate experiential drives only (though this would later be recognized as not a completely accurate account through the findings of EXPs 2+3). It appeared that in EXP.1 participants were altering performance as a function of their previous musical training, where Musicians were found to invert the lowered-then-raised pitch manipulation while non-musicians showed no reliable effect of the treatment. Importantly, the degree of divergence over time for musicians (but only the musicians) was also found to follow the proportion of manipulation, despite the inversion of this effect. Musicians becoming less like the background noise seemed reasonable for a number of possibilities. For example, recognizing a potential role of experience and expectation, musicians are well-practiced communicating against ambient music, and therefore it seemed possible they may have adopted strategies to compensate for such noise. If global pitch of the ambient noise was lowered, then raising voice-pitch toward areas in the spectrum which are more scant in their representation should equate to relatively less noise masking. This type of strategy has been noted previously in a host of avian research, where multiple species of birds living in urban areas have been found to exhibit relatively higher mean pitch than the same species living in rural areas (e.g., Slabbekoorn, 2013; Nemeth & Brumm, 2010; Halfwerk & Slabbkoorn, 2009; Ríos-Chelén et al., 2012). This effect has been described as bird-vocalizations diverging from the

anthropogenic noise (that is, noise from traffic, construction, etc.) in the environment, and has been argued to improve communicative abilities and potentially aid in reproductive fitness.

However, in retrospect, the findings of EXPs 2+3 suggest differences in pitch were in fact not rooted in effective message transmission, but instead are a reflection of speaker attitudes and opinions – though, such socially-driven effects were (at times) further mediated by a speaker's listening habits. That is to say, EXPs 2+3 show that how much music a participant listens to daily was a significant predictor for speaker convergence and divergence; though, the patterning of altered speech suggests that people who listen to relatively more music were converging (much like speakers who are known to converge with speech partners they hold high opinions of) and those who listen to relatively less music were found to diverge (much like speakers dissociating themselves from speakers or speech groups they do not ally with). Looking at distributions of the data showed that the musicians in EXP.1 reported listening to relatively little music, and the non-musicians reported listening to more music per day. With the exception of the intensity-based manipulation (discussed above), it appears these findings are consistent across studies 1 and 2. Therefore, both social and experiential motivations for convergence and divergence have been supported by the data in EXP.1, though a post-hoc exploration of participant distributions in these early data was required in order to recognize that explanations from later studies also hold within EXP.1 data.

Before moving on to the second experiment, attention should also be drawn to novel methods of stimuli generation applied in this work; recall that this process involved software previously unknown to phonetic enquiry. In addition to supplying evidence that speech production can be influenced by background noise, experiment 1 has also introduced and tested a novel means of stimuli generation and manipulation through Ableton's Live 9. Though proprietary in nature, this programme has been confirmed to achieve good results with minimal effort in contexts where other software failed to achieve the desired acoustic manipulations. For these reasons, Live 9 seems likely a valuable tool in future phonetic studies.

6.2.2 Experiment 2

Experiment 2 was designed primarily as a replication of EXP.1, aiming to maintain relative comparability while addressing the issues which spurred the previous low-power analysis. This study therefore refined the methods of EXP.1 through altered production stimuli (i.e., reading materials); global manipulations by Block (not progressive over time within a single condition); and the addition of a distractor task in the form of a video game, played in between treatments, to minimize the potential for effects that may persist after a condition ends (e.g., Brown, 1958). Making these changes and increasing participant numbers resulted in a more robust, straightforward analysis unencumbered by the issues of time series data.

As mentioned during the immediately preceding summary (6.2.1), trends largely similar to those observed in EXP.1 were also recognized in the data collected during EXP.2 – however, these higher-power, more complex analyses suggested musical training alone was insufficient to predict these effects. Instead, a self-reported estimate of HoursMusic listened to per day (on average) is implicated as the primary predictor, where speakers who listen to 1-hour of music a day or less appear to diverge for the most part (cf. EXP.1 Musicians) and those who report listening

to more than one hour of music per day are found to converge with the acoustic manipulations (cf. EXP.1 Non-musicians). Importantly, musicianship retained significance in certain interactions and was recognized to further shape the realizations of these convergent/divergent effects. That is, while convergent effects were observed for all degrees of musicianship in both the pitch and intensity conditions, the musicians and non-musicians did not exhibit the predicted divergence during the intensity-based treatment (though, the SMs did, and all degrees of musicianship were found to diverge during the pitch-based treatment). Thus, findings appear quite similar across studies, though more fleshed out during the EXP.2 analysis. The fact that these effects are observed and seem consistent across experiments serves as stronger evidence that background noise does, indeed, influence the fine-acoustic detail of a speaker's productions in predictable ways.

Another important outcome from EXP.2 can be recognized specifically through altered performance in response to the intensity manipulation. It was unclear through EXP.1 whether or not intensity-based changes observed were indeed effects of convergence, or if they would better have been attributed to the Lombard effect. In fact, it was unclear whether or not these processes are distinct. One way to address these questions would be to present signals below the known threshold for eliciting Lombard speech. To this end, intensity-levels in EXP.2 were manipulated to drop 6 dB below the known threshold for eliciting Lombard speech (Lazarus, 1986), where significant effects of both convergence and divergence were observed in EXP.2 (outcomes dependent upon the speaker's listening habits). While it is possible that the known Lombard-threshold is actually lower than previously believed, both the complementary performance observed across pitch- and intensity-based conditions (involving both convergence and divergence), and the fact that convergence was observed at a presentation level well-below the known Lombard threshold supports entrainment and the Lombard effect as distinct. Moreover, the lack of divergence observed in EXP.1 further suggests those intensity-based effects were most likely driven by the Lombard response, where as effects of convergence/divergence observed in EXP.2 (where manipulations fall below the known threshold) were instead driven by a different process.

Finally, when considering the second research question, the claim that some of the variation observed in these data was driven by socially-based motivations still holds. Indeed, it still seems the case that some experiential based motivations are also contributing to the realization of these convergent and divergent effects. As described in section 6.2.1, we see that the majority of speakers who listen to relatively little music per day tend to diverge from the acoustic manipulations in the background music while speakers who listen to relatively more music are found to converge with acoustic characteristics of the background music. While these effects parallel socially-driven findings in the sociolinguistic literature, we also see that these effects are mediated, though in a limited capacity, by experiential drives – that is, a speaker's level of musical training was predictive of negated and inverted effects in the intensity-based condition, though only for speakers who report listening to one hour or less of music per day. Between studies one and two, though, both social and experiential drives have been supported for acoustic convergence and divergence to background noise.

6.2.3 Experiment 3

Having shown intensity-based effects both above and below the known threshold for eliciting a Lombard response, and thus providing reasonable evidence for these processes as distinct, Experiment 3 focuses solely on entrainment observed in voice-pitch. Methods are identical to those of EXP.2, with the exception of a new speech-based condition; this experiment tests for entrainment in voice-pitch to background music and background speech, and has been designed to do so in a way where effects across conditions are directly comparable. Not assuming these processes to work necessarily in the same way, models were fit separately to speech- and music-based data.

Through this study I have supplied further evidence supporting a reliable and predictable influence of background noise on human speech production, where speech becomes systematically more and less like ambient noise under certain conditions. However, models in EXP.3 indicate predictor structures that differ once again from those of EXPs 1+2, and which also vary by noise type (Table 5.3). With regard to speech-based effects, we find only an influence of identified gender predicting convergent behaviours, where female speakers show no effect and male speakers are observed to align with the lowered-pitch speech signals. While this might seem at first a strange division, gender-based splits in effects of convergence are not uncommon in sociolinguistic research (e.g., Hay & Drager, 2010; Drager, Hay & Walker 2010).

I have argued earlier that I know of no physiological differences between male and female speaker which might account for any of the differences in behaviour observed in this work; therefore, I regard any such gender-based differences as largely experiential in nature, and most likely attributable to different experiences one might gain through living and learning as a female vs. as a male. In fact, Babel (2009) describes a similar split in convergence-based effects involving both gender and attractiveness ratings, where female speakers were found to reliably converge with speakers rated as more attractive while male speakers were found to diverge from them – it seems likely that in both Babel's data and in the present work, that personal experience is shaping the social motivations where both contribute to effects of convergence and divergence. Thus, the behaviours observed within the present work are motivated by both theory and previous research to be driven by participant attitudes and alliances, as well as specific forms of personal experience (recognizing that these two motivations can be somewhat related at times).

With regard to music-based effects of convergence/divergence in these data, models indicate that patterns of convergence and divergence were best explained by the ChooseMusic variable (that is, whether or not one chooses to listen to music while undertaking cognitively demanding tasks); however, these effects were further modulated by a speaker's identified gender. Through the interaction we find that females who ChooseMusic are predicted to converge acoustically, and men who do not ChooseMusic show no significant effect. Conversely, females who do not ChooseMusic show no significant effect of the manipulation, whereas the males who do not ChooseMusic during cognitively demanding tasks are found to invert this effect (i.e., diverge). There is no clear reasoning by which IdentifiedGender should counter the effect of ChooseMusic in this context, though the primary realization of this ChooseMusic effect seems much like the HoursMusic variable implicated in EXP.2. Specifically, looking at Figure 4.12 indicates an intuitive relationship: Participants who do not choose to listen to music during cognitively demanding tasks also tend to listen to less music generally, while participants who *do* choose to listen to music

during cognitively demanding tasks also generally report listening to more music per day on average. Therefore, while certain differences in performance driven by IdentGender are not easily interpreted (i.e., null effects), the general trends observed as a function of *musical consumption* (or, listening habits) are very much like those observed throughout EXPs 1+2, therefore further supporting speakers' productions as reliably becoming more (and sometimes less) like backgrounded music. Furthermore, as expected, music-based predictors were *not* recognized as significant in the speech-based mixed effects models, therefore reinforcing the validity of these effects in the music-based treatments. Once again, this study provides coherent evidence that fits well enough with known theory, where both social and experiential motivations appear to shape convergent and divergent behaviour in speech production.

As a final point of note before moving on, it is interesting that *any* effects were observed in response to these speech-based treatments in light of Lazarus' (1986) description of the known threshold for eliciting Lombard speech through speech-based signals. If Lazarus was correct identifying this lower limit as 55 dB for ambient speech, then the effects observed presently may serve as further evidence that entrainment and the Lombard effect are distinct processes. Admittedly, these are not directly comparable situations when contrasting pitch vs. intensity-based manipulations; though, recognizing that spectral changes (including altered F0) also take place in Lombard responses, altered productions at such a low presentation level become remarkable.

6.2.4 A Unified Analysis

Because the effects of convergence and divergence (related to music-based stimuli) were attributed to another new predictor in EXP.3, it seemed reasonable to attempt a unified analysis to test whether or not data from multiple experiments might be explained through a single predictor (or interaction). Specifically, while unlikely, it seemed possible that the observed effects could have been erroneously recognized as patterns in the data, where any spurious effects should likely not withstand the expanded dataset of a unified analysis. Therefore, data were transformed in such a way as to make observations collected during EXPs 1-3 relatively more comparable. However, while not unexpected given the differences in experimental design and data collection, observations from EXP.1 were largely uninterpretable in this context and were subsequently removed from the dataset for adding excessive noise. Data from EXPs 2+3 were designed to be directly comparable though, and resulted in a relatively coherent analysis.

This analysis confirms effects observed in earlier analyses can withstand a more robust dataset, where IdentifiedGender continues to interact significantly with BinaryHours in a way that modulates this effect, despite the inclusion of more data. Through combined datasets, we find that female speakers who report listening to one hour or less of music per day tend to exhibit significant effects of divergent behaviour across studies, while female speakers who report listening to more than one hour per day exhibit significant convergent behaviour. The male speakers, however, exhibit no significant effects of convergence or divergence. While the effects of IdentGender are once again difficult to interpret, though again appear to negate some of the effects associated with musical listening habits, the performance of female speakers retains patterns observed in previous analyses. Thus, this unified analysis supports both the social and experiential motivations described in previous chapters as genuine, and not the product of Type I errors.

6.3 Theory, Implications, and Addressing Research Questions

In light of the analyses discussed above, the evidence for convergence and divergence to acoustic characteristics of background noise appears relatively solid. Returning, then, to the research questions that drive the present work, **(1) Are speakers reliably influenced by acoustic characteristics of ambient noise, such that acoustic characteristics of their speech become systematically more or less like corresponding characteristics of that noise?** *It does appear that speakers are reliably influenced by acoustic characteristics of ambient noise, and that aspects of their speech become systematically more (and sometimes less) like that noise*. We have seen convergent and divergent behaviour in speech production across studies and across noise types, where separate analyses consistently identify predictors to explain significant differences in speakers' productions observed by treatment. While variables with relatively increased predictive power may be identified in the future, which importantly are likely to vary by noise type, musical consumption habits and identified gender appear to maintain reasonable explanatory power throughout the present work – though, to varying degrees across studies. Indeed, examining distributions of these variables across experimental datasets reinforces the notion that there are some underlying characteristics shared by speakers who diverge, and by those who converge.

The evidence thus far suggests that these effects are driven in part by something akin to 'practiced listening' where, by virtue of higher levels of music-consumption, these speakers are increasingly susceptible to convergent behaviour. But why would this be? As mentioned above, previous linguistic-research has shown that speakers tend to converge with those they identify with or hold positive opinions of, while they tend to diverge from speakers whom they do not identify with or hold negative opinions of. The effects of convergence/divergence observed in the present work may be explained by similar reasoning, where it appears the function in this context may be related in purpose. That is to say, where speakers converging to other speakers often signifies solidarity, and therefore is serving a largely social function, it seems possible that speakers who like music (and listen to more of it) alter their productions to become more like background music as a reflection of their attitudes. That is to say, perhaps convergence in this context is minimizing a social distance, and therefore expresses a fondness (or, an alliance of sorts) with something the speaker enjoys. On the other hand, speakers are known to diverge linguistically from other speakers whom they wish to distance themselves from for one reason or another. In much the same way, it appears possible that speakers with a distaste for music, or at least those who choose to listen to relatively little of it, also alter speech to become less like the background music as a reflection of their attitudes, thereby maximizing a social distance and expressing an aversion to something they are not fond of.

This line of thought serves as a return to the second research question driving the present work: **(2) Is there meaningful variation between individuals or groups in terms of convergent and divergent behaviour – and, if so, to what extent might this variation be predictable through social- or experience-based variables?** *The short answer to this question appears to be "yes."* Much like Babel (2009) argues, these data too support purely social and purely automatic theories of convergence as too strong. Operationally defining *automatic* in this context to reflect arguments made in Babel (2010) – that is, meaning specifically that convergent/divergent effects are automatic insofar as speakers are most often unaware that these changes take place, while these processes are simultaneously

not automatic in that they do not always take place in all contexts and instances (note: I will use the term *automated* below to refer to automatic processes that are more regular and necessarily occur) – we certainly see hints of automaticity in much of the observed effects, as speakers reported being unaware of the changes they made in these convergent/divergent contexts. In fact, it appears that we also see convergence and divergence in speech production *automatically* reflecting speaker attitudes about music throughout the three experiments. Where it has been argued by Trudgill (2008) that social affinities do not govern convergence/divergence but instead are fostered as products of accommodation, this argument does not appear to satisfactorily explain the data collected during the present work. It cannot be the case that speakers converge/diverge acoustically with background music in a completely automated way and then begin to form opinions about music after the fact, if for no other reason than the descriptions of musical consumption were gathered at the same time that acoustic data were collected in all three experiments (that is, within a single session). Thus, music-related opinions (reflected through consumption habits) must have existed beforehand. As a result, the experiments described in this dissertation serve to further support acoustic convergence most often as a reflection of attitudes, and expressly not the driving force behind them.

However, aiming to better understand the nature of these automatic effects, I want to leave the reader with a question that might be explored through future work regarding two possible functions/drives for acoustic convergence and divergence. Specifically, it seems possible, as I have mentioned above, that the examples of convergence and divergence described within this work are covert expressions of speaker-attitudes – though, I raise the possibility that perhaps convergence and divergence are instead only bi-products of these attitudes. The difference between these two scenarios is intention: Are speakers unconsciously aiming to express opinions and alignments through these changes in speech production, or are such changes instead only reflexive responses to the object of convergence or divergence? In the first instance these processes are purposeful, where the speaker is signalling important information through these changes. In the second, by analogy, convergence and divergence are effectively exhaust from a car. Sociolinguistic theory has suggested the former (i.e., signalling) is the case with regard to convergence in the context of communicative speech. The effects described within the present work, however, appear very much like those observed in previous accommodation efforts in their realization – where the likelihood of speakers aiming to align themselves socially with a musical composition seems much less likely the case. Because the realization of these effects across contexts (communicative vs. ambient) is rather similar, our knowledge about these processes, as well as related accommodation theory, would benefit from making this distinction. Acoustic accommodation in the context of communicative speech can still be meant to signal speaker attitudes even if convergence to ambient music does not; however, this type of purposeful availability of information is not necessarily something that should be taken for granted.

Following mention of effects that may be driven by *practiced listening*, one further line of thought involves the potential for an alternative interpretation of certain "divergent" behaviours observed throughout this work. As opposed to signalling, much of the divergence might also be explained through the Lombard effect. What if those who listen to more music converge because they can actually tune-in to information otherwise unavailable (or are more susceptible to its influence) – and those who listen to less/no music are exhibiting a Lombard response and *not really diverging*? This interpretation could explain divergent pitch in both EXPs 2+3 (rising F0) as well as the

divergent intensity observed in the same studies. This alternative interpretation could also potentially explain convergent behavior to intensity and divergent pitch observed in EXP.1 where a Lombard response could potentially overpower effects of convergence (as I've argued might be the case earlier). The potential for an explanation for apparent divergence through a Lombard response, however, does not account for non-musicians converging to lowered intensity in EXP.2, nor does it explain a lack of increased voice-pitch in the B/C comparison during the pitch-based manipulation in EXP.1. For these reasons, the explanation based in divergence has been accepted as most parsimonious; though, further study would be required to conclusively distinguish these possibilities, where experimental design would play a crucial role in recognizing the drive for this behavior.

As a final thought, with regard to speech-based conditions and the corresponding variation observed in production, it is worth noting that the present work has tested (and found effects) for entrainment to extra-linguistic (i.e., prosodic) information, where the only previous work (Delvaux & Soquet, 2007a; 2007b) explored and found effects for phonemic convergence. If no effects of convergence to background speech had been observed in the present work, this would have suggested speakers entrain to ambient linguistic information within the signal and not to prosodic information. If speakers converged with speech only (and not music), this could have suggested there is something special about speech processing that distinguishes linguistically communicative sounds from all other auditory input (e.g., Liberman, Cooper, Harris & Macneilage, 1962; Liberman & Mattingly, 1985; Fowler, 1989; McGurk & MacDonald, 1976), or perhaps that there is a required social drive motivating all forms of accommodation. It would also have been possible in the present work that speakers could have converged only to music (and not speech, reinforcing issues noted regarding Delvaux & Soquet, 2007a; 2007b), which would suggest once again that ambient speech is processed differently than other forms of noise, and that human listeners/talkers are perhaps better equipped to dismiss the multi-talker signals as distracting noise. However, the current findings indicate that none of these are the case. We have seen effects of convergence and divergence to both background music and multi-talker speech, though analyses suggest these processes are somewhat different in their nature.

In summary, through this dissertation I have shown that speakers are often influenced in speech production by ambient noise, and that productions often become systematically more (and sometimes less) like that noise. I have shown that these effects appear to be related to both a speaker's identified gender and their musical listening habits, where speakers who listen to relatively less music often trend toward divergence whereas speakers who listen to relatively more music most often trend toward convergence. Indeed, these effects merge well with the current knowledge of accommodation in communicative speech which argues speakers are more likely to converge with speakers they enjoy on some important level, and diverge from those they do not. And, although I have found effects of convergence/divergence to both ambient music and to ambient speech, analyses within this work suggest these processes are different in some important ways. Because human speakers are nearly always contending with some form(s) of ambient noise though, the present work suggests that we may be constantly updating speech production patterns to become more and less like that environmental noise at different times. Therefore, this and later works seem very likely to play an important role in spurring the development of speech production and processing theories in the future.

CHAPTER 7: References

Ableton (2015). Ableton Live (Version 9) [Computer program]. retrieved from http://www.ableton.com

Adank, P., Hagoort, P., & Bekkering, H. (2010). Imitation improves language comprehension. *Psychological Science*, *21*(12), 1903-1909.

Amazi, D. K., & Garber, S. R. (1982). The Lombard sign as a function of age and task. *Journal of Speech, Language, and Hearing Research*, *25*(4), 581-585.

Apple Inc. (2012). GarageBand [Computer program]. Version 6.0.5, retrieved from http://www.apple.com/

Aylett, M., and Turk, A. (2004). "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," Lang. Speech 47, 31–56.

Aylett, M., and Turk, A. (2006). "Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei," J. Acoust. Soc. Am. 119(5), 3048–3058.

Baayen (2013). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 1.4.1.

Baayen, R., Piepenbrock, R. & Gulikers, L. (1995). CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium.

Baayen, R. H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics Using R (Cambridge University Press, New York).

Baayen, R.H. (2013). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics".. R package version 1.4.1. https://CRAN.R-project.org/package=languageR

Babel, Molly. (2010). Dialect convergence and divergence in New Zealand English. *Language in Society*, 39 (4), 437-456.

Babel, M. (2009). Phonetic and social selectivity in speech accommodation. Doctoral dissertation, University of California, Berkeley. Retrieved from: https://escholarship.org/uc/item/1mb4n1mv

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40(1), 177-189.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.

Bauer, J. J., Mittal, J., Larson, C. R., & Hain, T. C. (2006). Vocal responses to unanticipated perturbations in voice loudness feedback: An automatic mechanism for stabilizing voice amplitude. *The Journal of the Acoustical Society of America*, *119*(4), 2363-2371.

Beckner, C., Rácz, P., Hay, J., Brandstetter, J., & Bartneck, C. (2016). Participants conform to humans but not to humanoid robots in an English past tense formation task. *Journal of Language and Social Psychology*, *35*(2), 158-179.

Bell, A. (1984). Language style as audience design. *Language in society*, *13*(2), 145-204.

Berlin, C.I., Lowe-Bell, S.S., Cullen, J.K., Jr., Thompson, C.L., and Loovis, C.F. (1973). Dichotic speech perception: An interpretation of right-ear advantage and temporal offset effects. *Journal of the Acoustical Society of America*. 53, 699-709.

Bilous, F. R., & Krauss, R. M. (1988). Dominance and accommodation in the conversational behaviours of same-and mixed-gender dyads. *Language & Communication*.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer (Version 5.4.04) [Computer program]. Retrieved from www.praat.org.

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. In D. Birdsong (Ed.) *Second language acquisition and the critical period hypothesis*, 133-159. Mahwah, NJ: Erlbaum.

Branigan, H. P., Pickering, M. J., McLean, J. F., & Cleland, A. A. (2007). Syntactic alignment and participant role in dialogue. *Cognition*, *104*(2), 163-197.

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, *96*, 41-44.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482.

Briefer, E. F., & McElligott, A. G. (2012). Social effects on vocal ontogeny in an ungulate, the goat, Capra hircus. *Animal Behaviour*, *83*(4), 991-1000.

Broadbent, D.E., (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*. 47, 191-196.

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465-1487.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12-21.

Brumm, H., & Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour*, 148(11-13), 1173-1198.

Caelen-Haumont, G. (2009). Emotion, emotions and prosodic structure: an analysis of the melisms patterns and statistical results in the spontaneous discourse of 4 female speakers from four generations. In Sylvie Hancil (Ed.) *The role of prosody in affective speech* (Vol. 97). Oxford: Peter Lang (pp. 95-138).

Candiria & Coma, C., (2001). Without Water (recorded by Candiria). On *300 Percent Density*. Hawthorne, CA: Century Media.

Chambers, J. K. (2002). Dynamics of dialect convergence. *Journal of Sociolinguistics*, *6*(1), 117-130.

Chen, J. L., Penhune, V. B., & Zatorre, R. J. (2008). Moving on time: brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *Journal of cognitive neuroscience*, *20*(2), 226-239.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, *25*(5), 975-979.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562-1573.

Cooke, M., & Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *The Journal of the Acoustical Society of America*, 128(4), 2059-2069.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111*(6), 2862-2873.

Currie Hall, K., Hume, B., Jaeger, F. T., & Wedel, A. (Under Review). The message shapes phonology. Downloaded 10 Janauary, 2018 from: https://www.researchgate.net/publication/309033386_The_Message_Shapes_Phonology

Darwin, C. J. (2007). Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1493), 1011-1021.

Delvaux, V., & Soquet, A. (2007a). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, *64*(2-3), 145-173.

Delvaux, V., & Soquet, A. (2007b). Inducing imitative phonetic variation in the laboratory. In *Proceedings ICPhS* (pp. 369-372).

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*(2), 163-185.

DeSantis, D., Gallagher, I., Haywood, K., Knudsen, R., Behles, G., Rang, J., Henke, R., & Slama, T. (2013). *Ableton Reference Manual Version 9: For Windows and Mac OS*. Berlin. Retrieved from: https://cdn2-resources.ableton.com/80bA26cPQ1hEJDFjpUKntxfqdmG3ZykO/static/manual/pdf/L9Manual_EN.pdf

Dilts, P. C. (2013). *Modelling phonetic reduction in a corpus of spoken English using Random Forests and Mixed-Effects Regression*. Doctoral Dissertation, University of Alberta. Retrieved from: https://era.library.ualberta.ca/items/ff36bb82-fe70-4a83-b244-96babc2a36bc/download/a03ff422-f3a6-4fe6-9732-b9ac19652a99

Down, S. (2009). *The effect of tempo of background music on duration of stay and spending in a bar*. Master's Thesis, University of Jyväskylä. Retrieved from: https://jyx.jyu.fi/bitstream/handle/123456789/20304/1/URN_NBN_fi_jyu-200905271640.pdf

Draegert, G. L. (1951). Relationships between voice variables and speech intelligibility in high level noise. *Communications Monographs*, 18(4), 272-278.

Drager, K., Hay, J., & Walker, A. (2010). Pronounced rivalries: Attitudes and speech production. *Te Reo*, *53*, 27.

Drager, K. K. (2011). Sociophonetic variation and the lemma. *Journal of Phonetics*, *4*(39), 694-707.

Drager, K., & Kirtley, J. (2016). Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. *Awareness and control in sociolinguistic research*, 1-24.

Drake, C. (1998). Psychological processes involved in the temporal organization of complex auditory sequences: Universal and acquired processes. *Music Perception: An Interdisciplinary Journal*, *16*(1), 11-26.

Duke, R. A. (1994). When tempo changes rhythm: the effect of tempo on nonmusicians' perception of rhythm. *Journal of Research in Music Education*, *42*(1), 27-35.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453-476.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, *47*(2), 421-436.

Egan, J. J., 1972. Psychoacoustics of the Lombard voice response. *Journal of Auditory Research*, 12 (4), 318–324.

Farnsworth, P. R., Block, H. A., & Waterman, W. C. (1934). Absolute tempo. *The Journal of General Psychology*, *10*(1), 230-233.

Farnsworth, P. R. (1950). *Musical taste: Its measurement and cultural nature*. Stanford, CA: Stanford University Press.

Flege, J. E., & Hammond, R. M. (1982). Mimicry of non-distinctive phonetic differences between language varieties. *Studies in Second Language Acquisition*, *5*(1), 1-17.

Fletcher, H., Raff, G. M., & Parmley, F. (1918). Study of the effects of different sidetones in the telephone set. *Western Electrical Company, Report* 19412.

Fowler, C. A. (1989). Real Objects of Speech Perception: A Commentary on Diehl and Kluender, *Ecological Psychology*, 1(2), 145-160.

Fraisse, P. (1982). Rhythm and tempo. In *The Psychology of Music*, D. Deutsch (Ed.). New York: Academic Press. pp.149-180.

Fromont, R., & Hay, J. (2008). ONZE Miner: the development of a browser-based research tool. *Corpora*, *3*(2), 173-193.

Galambos, R., Makeig, S., & Talmachoff, P. J. (1981). A 40-Hz auditory potential recorded from the human scalp. *Proceedings of the national academy of sciences*, *78*(4), 2643-2647.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*(3), 181-215.

Geva, E., & Ryan, E. B. (1993). Linguistic and cognitive correlates of academic skills in first and second languages. *Language learning*, *43*(1), 5-42.

Gibson, A. (2008). Perception of Sung and Spoken Vowels in New Zealand English. Presented at the 11th annual Laboratory Phonology (*LabPhon*) meeting.

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological linguistics*. 15, 87-105.

Giles, H., & Powesland, P. F. (1975). *Speech style and social evaluation*. Oxford, England: Academic Press.

Giles, H., Mulac, A., Bradac, J. J., & Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. *Annals of the International Communication Association*, *10*(1), 13-48.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*. New York: Cambridge University Press. 1- 68.

Giles H., Powesland P. (1997) Accommodation Theory. In: Coupland N., Jaworski A. (eds) Sociolinguistics. Modern Linguistics Series. Palgrave, London.

Gnevsheva, K. (2017). Within-speaker variation in passing for a native speaker. *International Journal of Bilingualism*, *21*(2), 213-227.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review*, *105*(2), 251-279.

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic bulletin & review*, *11*(4), 716-722.

Graham, C. A., & McGrew, W. C. (1980). Menstrual synchrony in female undergraduates living on a coeducational campus. *Psychoneuroendocrinology*, *5*(3), 245-252.

Gregory, S. W. (1983). A quantitative analysis of temporal symmetry in microsocial relations. *American Sociological Review*, 129-135.

Gregory, S. W. (1990). Analysis of fundamental frequency reveals covariation in interview partners' speech. *Journal of Nonverbal Behaviour*, *14*(4), 237-251.

Gregory, S. W., Dagan, K., & Webster, S. (1997). Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behaviour*, *21*(1), 23-43.

Halfwerk, W., & Slabbekoorn, H. (2009). A behavioural mechanism explaining noise-dependent frequency use in urban birdsong. *Animal behaviour*, *78*(6), 1301-1307.

Harlow, R., Keegan, P., King, J., Maclagan, M., & Watson, C. (2009). The changing sound of the Māori language. In Stanford, J., Preston, D. (Eds.) *Variation in indigenous minority languages*. John Benjamins, Amsterdam.

Härtel, S. (2009). *Quinn* [computer program]. Retrieved from http://download.cnet.com/Quinn/3000-2099_4-32916 .html

Harrell, F. E. Jr. (2016/2018). rms: Regression Modelling Strategies. R package version 4.5-0. https://CRAN.R-project.org/package=rms

Harrison, R. (1941). Personal tempo and the interrelationships of voluntary and maximal rates of movement. *The Journal of General Psychology*, *24*(2), 343-379.

Hay, J., Drager, K., & Warren, P. (2009). Careful who you talk to: An effect of experimenter identity on the production of the NEAR/SQUARE merger in New Zealand English. *Australian Journal of Linguistics*, *29*(2), 269-285.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, 48(4), 865-892.

Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *The linguistic review*, *23*(3), 351-379.

Hay, J., Podlubny, R., Drager, K., & McAuliffe, M. (2017). Car-talk: Location-specific speech production and perception. *Journal of Phonetics*, *65*, 94-109.

Hayes, B., & Lahiri, A. (1991). Bengali intonational phonology. *Natural Language & Linguistic Theory*, *9*(1), 47-96.

Helmholtz, Hermann. (1954). *On the Sensations of Tone*, translated by Alexander J. Ellis from the fourth German edition (1887). New York: Dover Publications.

Heinks-Maldonado, T. H., & Houde, J. F. (2005). Compensatory responses to brief perturbations of speech amplitude. *Acoustics Research Letters Online*, *6*(3), 131-137.

Herz, R. S., & Engen, T. (1996). Odor memory: Review and analysis. *Psychonomic Bulletin & Review*, *3*(3), 300-313.

Hill, A. R., Adams, J. M., Parker, B. E., & Rochester, D. F. (1988). Short-term entrainment of ventilation to the walking cycle in humans. *Journal of Applied Physiology*, *65*(2), 570-578.

Hirschberg, J. (1993). Pitch accent in context: predicting intonational prominence from text. *Artificial Intelligence*, *63*(1-2), 305-340.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. (2006). Survival Ensembles. Biostatistics, 7(3), 355--373.

Hovdhaugen, E. (1992). Phonetic vowel length in Samoan. *Oceanic Linguistics*, 31(2), 281-285.

Howell, P. (2008). Effect of speaking environment on speech production and perception. *Journal of the human-environment system,* 11(1): 51–57.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., & Johnson, K. (1999). Formants of children, women, and men: The effects of vocal intensity variation. *The Journal of the Acoustical Society of America*, *106*(3), 1532-1542.

Hugenii, C. (1986). Horoloquim oscilatorium. *Paris: Muguet. Reprinted in English as: The pendulum clock. Ames, IA: Iowa State UP*.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57-83.

Jenkins, J. (2000). *The Phonology of English as an International Language*. Oxford: Oxford University Press.

Junqua, J. C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication*, *20*(1), 13-22.

Kantono, K., Hamid, N., Shepherd, D., Yoo, M. J., Carr, B. T., & Grazioli, G. (2016). The effect of background music on food pleasantness ratings. *Psychology of Music*, *44*(5), 1111-1125.

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological science*, *23*(6), 661-668.

Kimura, D. (1964). Right-left differences differences in the perception of melody. *Quarterly Journal of Experimental Psychology*. 16, 355-358.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. D. (2004). Music, language and meaning: brain signatures of semantic processing. *Nature neuroscience*, *7*(3), 302-307.

Krainik, A., Lehericy, S., Duffau, H., Capelle, L., Chainay, H., Cornu, P., Cohen, L., Boch, A.-L, Mangin, J.-F, Le Bihan, D., & Marsault, C. (2003). Postoperative speech disorder after medial frontal surgery Role of the supplementary motor area. *Neurology*, *60*(4), 587-594.

Kunnari, S., Nakai, S., & Vihman, M. M. (2001). Cross-linguistic evidence for acquisition of geminates. *Psychology of Language and Communication*, *5*(2), 13-24.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech, Language, and Hearing Research*, *14*(4), 677-709.

Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behaviour*, *27*(3), 145-162.

Lazarus, H. (1986). Prediction of Verbal Communication is Noise—A review: Part 1. *Applied Acoustics*, *19*(6), 439-464.

LeBlanc, A., & McCrary, J. (1983). Effect of tempo on children's music preference. *Journal of Research in Music Education*, *31*(4), 283-294.

Lee, G., Lifeson, A., Peart, N. (1981). Limelight (recorded by Rush). On *Moving Pictures*. New York, NY: Mercury Records.

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In the proceedings of the *Twelfth Annual Conference of the International Speech Communication Association*, pp. 3081-3084.

Lieberman, P. (1967). *Intonation, perception, and language*. Cambridge: The MIT Press.

Liberman, A. M. (1984). On finding that speech is special. In *Handbook of Cognitive Neuroscience* (pp. 169-197). Springer US.

Liberman, A.M., Cooper, F.S., Harris, K.S., and Macneilage, P.F., (1962). A motor theory of speech perception, Presented at the Speech Communication Seminar, Stockholm, 1-12.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21(1). 1-36.*

Lindblom, B. (1990). "Explaining phonetic variation: A sketch of the H & H theory," in Speech Production and Speech Modeling, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic, Dordrecht, The Netherlands), pp. 403–439.

Lloyd, M., & Dybas, H. S. (1966). The periodical cicada problem. I. Population ecology. *Evolution*, *20*(2), 133-149.

Lombard, E. (1911). Le signe de l'elevation de la voix. Ann. *Maladies Oreille, Larynx, Nez, Pharynx*, 37(101-119), 25.

London, J. (2012). *Hearing in time: Psychological aspects of musical meter*. (2nd Edition). New York: Oxford University Press.

Longstreth, D. (2012). *Offspring are Blank* (recorded by the Dirty Projectors). On *Swing Low Magellan*. Brooklyn, NY: Domino Records.

Love, J., & Walker, A. (2013). Football versus football: Effect of topic on/r/realization in American and English sports fans. *Language and speech*, *56*(4), 443-460.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & psychophysics*, *62*(3), 615-625.

Manson, J. H., Bryant, G. A., Gervais, M. M., & Kline, M. A. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behaviour*, *34*(6), 419-426.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*(5417), 522.

Marx, N. (2002). Never quite a 'native speaker': Accent and identity in the L2-and the L1. *Canadian Modern Language Review*, *59*(2), 264-281.

McClintock, M. K. (1971). Menstrual synchrony and suppression. *Nature*, Vol 229, 244-245.

McGurk H., & MacDonald J. (1976). Hearing lips and seeing voices, *Nature*, 264, 746-748.

Meyer, L. B. (1957). Meaning in music and information theory. Journal of Aesthetics and Art Criticism, 15(4), 412–424.

Miller, G. A. (1947). The masking of speech. *Psychological bulletin*, *44*(2), 105.

Mirollo, R. E., & Strogatz, S. H. (1990). Synchronization of pulse-coupled biological oscillators. *SIAM Journal on Applied Mathematics*, *50*(6), 1645-1662.

Miyake, Y. (2009). Interpersonal synchronization of body motion and the Walk-Mate walking support robot. *IEEE Transactions on Robotics*, *25*(3), 638-644.

Moelants, D. (2002). Preferred tempo reconsidered. In *Proceedings of the 7th international conference on music perception and cognition* (pp. 580-583).

Moelants, D. (2003). Dance music, movement and tempo preferences. In *Proceedings of the 5th Triennal ESCOM Conference*.

Morrill, T. H., Dilley, L. C., & McAuley, J. D. (2014). Prosodic patterning in distal speech context: Effects of list intonation and F0 downtrend on perception of proximal prosodic structure. *Journal of Phonetics*, *46*, 68-85.

Murray-Smith, R., Ramsay, A., Garrod, S., Jackson, M., & Musizza, B. (2007). Gait alignment in mobile phone conversations. In *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services* (pp. 214-221). ACM.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*(4), 422-432.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, *32*(5), 790.

Nemeth, E., & Brumm, H. (2010). Birds and anthropogenic noise: are urban songs adaptive?. *The American Naturalist*, *176*(4), 465-475.

Odden, D. (1995). Tone: African languages. In J. A. Goldsmith (Ed.), *Handbook of phonological theory*. Oxford: Blackwell.

Pantaleone, J. (2002). Synchronization of metronomes. *American Journal of Physics*, *70*(10), 992-1000.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382-2393.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190-197.

Pardo, J. (2013). Measuring phonetic convergence in speech production. *Frontiers in psychology*, *4*, 559.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637-659.

Park, S. A. (2008). Consumer health crisis management: Apple's crisis responsibility for iPod-related hearing loss. *Public Relations Review*, *34*(4), 396-398.

Patel, A. D., & Peretz, I. (1997). Is music autonomous from language? A neuropsychological appraisal. In I. Deliège & J. Sloboda (Eds.), *Perception and cognition of music* (pp. 191-215). Hove, England: Psychology Press/Erlbaum (UK) Taylor & Francis.

Patel, A. D. (2008). *Music, language, and the brain*. New York: Oxford university press.

Patel, A. D., Iversen, J. R., Bregman, M. R., & Schulz, I. (2009). Avian and human movement to music: Two further parallels. *Communicative & Integrative Biology*, *2*(6), 485-488.

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, *50*(1), 302-313.

Pearce, M. T., & Wiggins, G. A. (2012). Auditory expectation: the information dynamics of music perception and cognition. *Topics in cognitive science*, *4*(4), 625-652.

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. *The neurocognition of language*, 167-208.

Pettit, M., & Vigor, J. (2015). Pheromones, feminism and the many lives of menstrual synchrony. *BioSocieties*, *10*(3), 271-294.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. Doctoral dissertation, Massachusetts Institute of Technology. Retrieved from:http://dspace.mit.edu/bitstream/handle/1721.1/16065/07492108.pdf?sequence=1

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. *Typological studies in language*, *45*, 137-158.

Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. New York: Springer Science & Business Media.

Platt, J., & Weber, H. (1984) Speech convergence miscarried: an investigation into inappropriate accommodation strategies. *International Journal of the Sociology of Language,* 46, 131-146.

Putman, W. B., & Street, R. L. (1984). The conception and perception of noncontent speech performance: Implications for speech-accommodation theory. *International Journal of the Sociology of Language*, (46), 97-114.

Psychology Software Tools. (2012). E Prime (Version 2.0) [Computer program].

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rao, M. D., & Letowski, T. (2006). Callsign acquisition test (CAT): Speech intelligibility in noise. *Ear and hearing*, *27*(2), 120-128.

Reichle, E. D., & Perfetti, C. A. (2003). Morphology in word identification: A word-experience model that accounts for morpheme frequency effects. *Scientific Studies of Reading*, *7*(3), 219-237.

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychonomic bulletin & review*, *12*(6), 969-992.

Repp, B. H., & Doggett, R. (2007). Tapping to a very slow beat: a comparison of musicians and nonmusicians. *Music Perception: An Interdisciplinary Journal*, *24*(4), 367-376.

Ríos-Chelén, A. A., Salaberria, C., Barbosa, I., Macías Garcia, C., & Gil, D. (2012). The learning advantage: bird species that learn their song show a tighter adjustment of song to noisy environments than those that do not learn. *Journal of evolutionary biology*, *25*(11), 2171-2180.

Roballey, T. C., McGreevy, C., Rongo, R. R., Schwantes, M. L., Steger, P. J., Wininger, M. A., & Gardner, E. B. (1985). The effect of music on eating behaviour. *Bulletin of the Psychonomic Society*, *23*(3), 221-222.

Russell, M. J., Switz, G. M., & Thompson, K. (1980). Olfactory influences on the human menstrual cycle. *Pharmacology Biochemistry and Behaviour*, *13*(5), 737-738.

Ryalls, B. O., & Pisoni, D. B. (1997). The effect of talker variability on word recognition in preschool children. *Developmental Psychology*, *33*(3), 441.

Ryan, W. J., & Burk, K. W. (1974). Perceptual and acoustic correlates of aging in the speech of males. *Journal of communication disorders*, *7*(2), 181-192.

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421-436.

Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5.

Schötz, S. (2007). Acoustic analysis of adult speaker age. In *Speaker Classification I* (pp. 88-107). Springer, Berlin, Heidelberg.

Schmuckler, M. A. (1989). Expectation in music: Investigation of melodic and harmonic processes. *Music Perception*, 7(2), 109-149.

Schulman, R. (1989). Articulatory dynamics of loud and normal speech. *The Journal of the Acoustical Society of America*, *85*(1), 295-312.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002) E-Prime User's Guide. Pittsburgh: Psychology Software Tools Inc.

Schweitzer, A., Lewandowski, N. (2013) Convergence of articulation rate in spontaneous speech. *In: Proc. of Interspeech*. pp. 525–529.

Shaw, P., & McMillion, A. (2008). Proficiency effects and compensation in advanced second-language reading. *Nordic Journal of English Studies*, *7*(3), 123-143.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422-429.

Siever, D. & Collura, T. (2017) "Audio–Visual Entrainment: Physiological Mechanisms and Clinical Outcomes" in *Rhythmic Stimulation Procedures in Neuromodulation*. Evans, J. R., & Turner, R. (Eds.). San Diego: Academic Press, 51-95.

Slabbekoorn, H., & Ripmeester, E. A. P. (2008). Birdsong and anthropogenic noise: implications and applications for conservation. *Molecular ecology*, *17*(1), 72-83.

Slabbekoorn, H. (2013). Songs of the city: noise-dependent spectral plasticity in the acoustic phenotype of urban birds. *Animal Behaviour*, *85*(5), 1089-1099.

Smith, P. C., & Curnow, R. (1966). " Arousal hypothesis" and the effects of music on purchasing behaviour. *Journal of Applied Psychology*, *50*(3), 255.

Smoll, F. L. (1975). Preferred tempo in performance of repetitive movements. *Perceptual and Motor skills*, *40*(2), 439-442.

Smoll, F. L., & Schutz, R. W. (1978). Relationships among measures of preferred tempo and motor rhythm. *Perceptual and Motor Skills*, *46*(3), 883-894.

Stern, K., & McClintock, M. K. (1998). Regulation of ovulation by human pheromones. *Nature*, *392*(6672), 177.

Stowe, L. M., & Golob, E. J. (2013). Evidence that the Lombard effect is frequency-specific in humans. *The Journal of the Acoustical Society of America*, *134*(1), 640-647.

Street Jr, R. L. (1984). Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, *11*(2), 139-169.

Strobl, C., Boulesteix, A.L., Zeileis, A., & Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics, 8(25). URL http://www.bio medcentral.com/1471-2105/8/25.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, *9*(1), 307.

Strobl, C., Hothorn, T., & Zeileis, A. (2009). Party on! – A New, Conditional Variable Importance Measure for Random Forests Available in the party Package. *The R Journal*, 1(2), 14–17.

Strogatz, S. H. (2000). From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, *143*(1-4), 1-20.

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language variation and change*, *24*(2), 135-178.

Taguchi, S., Gliner, J. A., Horvath, S. M., & Nakamura, E. (1981). Preferred tempo, work intensity, and mechanical efficiency. *Perceptual and motor skills*, *52*(2), 443-451.

Tam, A. (2017). *Neuromuscular Control of Vocal Loudness in Adults as a Function of Cue*. Masters dissertation, University of Alberta. Retrieved from: https://era.library.ualberta.ca/files/c2514nk86z/Tam_Andrea_J_201705_MSc.pdf

Trainor, L. (2008). Science & music: the neural roots of music. *Nature*, *453*(7195), 598.

Trudgill, P. (2008). Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation. *Language in Society*, *37*(2), 241-254.

Tweedy, R. S., & Culling, J. F. (2014). Does the signal-to-noise ratio of an interlocutor influence a speaker's vocal intensity?. Computer Speech & Language, 28(2), 572-579.

van der Wege, M. M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, *60*(4), 448-463.

van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, *84*(3), 917-928.

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision*, *8*(9), 14-14.

Walker, T. J. (1969). Acoustic synchrony: two mechanisms in the snowy tree cricket. *Science*, *166*(3907), 891-894.

Wang, W. (1967). Phonological features of tone. *International Journal of American Linguistics*, 33(2), 93-105.

Warren, S. (2001). Phonological acquisition and ambient language: A corpus based cross-linguistic exploration. Doctoral dissertation, University of Hertfordshire. Retrieved from: https://uhra.herts.ac.uk/handle/2299/14157.

Warner, N. (2012). Methods for studying spontaneous speech. Chapter in A. Cohn, C. Fougeron, & M. Huffman (eds.), *The Oxford Handbook of Laboratory Phonology*. Oxford: Oxford University Press, 621-633.

Webb, J. T. (1972). Interview synchrony: An investigation of two speech rate measures in an automated standardized interview. Chapter in A.W. Siegman, & B. Pope (eds.), *Studies in Dyadic Communication*. New York: Pergamon Press, 115-133.

Weissenböck, N. M., Schwammer, H. M., & Ruf, T. (2009). Estrous synchrony in a group of African elephants (Loxodonta africana) under human care. *Animal reproduction science*, *113*(1-4), 322-327.

Will, U., & Berg, E. (2007). Brain wave synchronization and entrainment to periodic acoustic stimuli. *Neuroscience letters*, *424*(1), 55-60.

Willemyns, M., Gallois, C., Callan, V. J., & Pittam, J. (1997). Accent accommodation in the job interview: Impact of interviewer accent and gender. *Journal of Language and Social Psychology*, *16*(1), 3-22.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73(1):3-36.

Yazan, B. (2015). Intelligibility. *ELT Journal*. 69(2), 202–204.

Yip, M. J. (1980). *The tonal phonology of Chinese*. Doctoral dissertation, Massachusetts Institute of Technology. Retrieved from: http://dspace.mit.edu/handle/1721.1/15971

Zorn, J., & Naked City. (1990). *You will be Shot* (recorded by John Zorn). On *Naked City*. New York, NY: Nonesuch Records.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 920.

CHAPTER 8: Appendices

APPENDIX 1 (Language Background Survey)

1. Age (in years) ?

2. What is the highest level of education you have completed (for example, 'completed 2nd year of high school' or 'enrolled at university now, have completed 2 years')?

3. What level of education did your father complete? What level of education did your mother complete?

4. Which hand do you use to write with?

5. Do/did you have a *hearing* impairment, or *reading* difficulties (e.g., difficulties learning to read), or *language development or speaking* difficulties (e.g., delayed language development, stuttering, lisping, etc.)?

6. What is your native language, i.e., the first language you learned, from birth? What is your father's native language? What is your mother's native language?

7. Please tell us every place you have lived for at least three months or more, starting with your infancy and childhood. Please be specific: if you lived in a city, please also indicate which area or town/suburb.

8. Were there any other people from a different city/town/region, or from a different country, who lived at your house or frequently spent time with you during your infant/childhood years (for example, grandmother watched over you while your parents were at work, a housekeeper or nanny)?

If so, who were they, and what country/city/town were they from, and what was their native language? Please also indicate what ages you were when they were often around?

9. Please tell us what *other* languages you speak besides English, and when and how long you have studied/learned each language, and how well you speak each one.

10. Have you recently returned from a long period in another country or in another city in New Zealand, longer than 3 months? If so, which country/city did you visit, and when did you return to Christchurch?

APPENDIX 2 (Data extraction Praat script function summaries)

1. 'Generate a textgrid for each recording/condition': Two nearly identical versions of this script were written. One was created for duration-altered stimuli and the other for all remaining stimuli, as bar lines are consistent for the duration of Science Music outside of temporal manipulation. Selection of which version of the script would be used took place (manually) prior to running the script. Therefore, the only difference between V1 and V2 was which pre-existing textgrid would be used for later concatenation.

   This script reads in all sound files within a specified directory, and then – file by file – identifies the onset of musical noise. The introduction of music is recognized automatically when intensity exceeds a preset threshold (using Praat's *To textgrid (silences)…* function). The time point where background music begins is used to align the appropriate base textgrid to that time point; then, from that point, Praat is instructed to calculate how much time precedes the onset of music for each file, and then generate a second textgrid where bar lines/intervals are back-filled within the appropriate tier every 1.854 seconds (the duration of a single bar in *Science Music*).[20] The two textgrids are then concatenated and automatically saved with a filename corresponding to the sound-file from which they were generated. As a failsafe, Praat is finally instructed to make sure the textgrid and sound file are of equivalent duration and pad time at the end of the textgrid if need be – this proves useful in instances where participants continue to read after the musical noise had stopped.
2. 'Get measures of interest': This script serves three main functions:
   a. All corresponding .wav/textgrid pairs from within a participant's folder are read into Praat. The left- (music/stimulus) and right-channels (speech recordings) are both extracted from the stereo recording. From the speech channel, pitch and intensity values are extracted at every pitch pulse. These values are printed to a text file, along with the time point at which each value was extracted and the corresponding bar number and section label, where *Section* was used to segment the stimuli on a gross level. Therefore, as seen below in Figure 1.7, Section-intervals indicate whether there was (a) No music, (b) Music with no manipulation, (c1) Music with the onset of manipulation, (c2) Music where manipulation reaches the maxima, and (c3) Music where manipulation returns to origin values.
   b. Histograms are printed by section in order to visualize the distribution of intensity measures by Section for each condition.
   c. A text file is generated as a comma separated value table for each participant, which includes descriptive statistics and labeling materials such as condition; max, min, mean, and standard deviation values by Section (for each condition); and bar numbers to reference each Section change.

[20] Because all conditions have been generated with the same consistent tempo before manipulation, intervals that precede the introduction of music should logically be segmented as the same 4-beat duration to maximize continuity and comparability.

APPENDIX 3 (EXP.1 Final model outputs)

Pitch - AB, Mus			
	Estimate	**Std.Error**	**t value**
(Intercept)	157.9216	23.1195	6.831
SectionB	-1.8351	0.2813	**-6.523**
Block	0.4546	0.2792	1.629
Pitch - AB, NoMus			
	Estimate	**Std.Error**	**t value**
(Intercept)	152.1024	16.0771	9.461
SectionB	-0.7406	0.2638	-2.807
Block	2.1443	0.1355	15.828
Pitch - BC, Mus			
	Estimate	**Std.Error**	**t value**
(Intercept)	160.654	22.9488	7.001
SecC	0.5297	0.2678	1.978
ConditionPitch	-2.7942	0.4363	**-6.404**
Block	-0.5392	0.1645	**-3.278**
SecC:ConditionPitch	1.3488	0.4613	**2.924**
Pitch - BC, NoMus			
	Estimate	**Std.Error**	**t value**
(Intercept)	156.75906	15.76086	9.946
SecC	-0.39758	0.24921	-1.595
ConditionPitch	-1.85941	0.39614	**-4.694**
Block	0.571	0.07313	**7.808**
SecC:ConditionPitch	0.46314	0.42405	1.092
Pitch - PropChange, Mus			
	Estimate	**Std.Error**	**t value**
(Intercept)	161.6821	22.9765	7.037
PropChange	-1.6924	0.3456	**-4.897**
ConditionPitch	-3.6899	0.3588	**-10.283**

Block	-0.4453	0.1776	**-2.507**
PropChange:ConditionPitch	4.4954	0.5983	**7.513**
Pitch - PropChange, NoMus			
	Estimate	**Std.Error**	**t value**
(Intercept)	157.06827	15.7534	9.97
PropChange	0.4491	0.32173	1.396
ConditionPitch	-1.56635	0.32277	**-4.853**
Block	0.26439	0.07867	**3.361**
PropChange:ConditionPitch	0.41192	0.54792	0.752
Intensity - AB, Mus			
	Estimate	**Std.Error**	**t value**
(Intercept)	59.7399	1.48443	40.24
SectionB	0.99537	0.03796	**26.22**
Block	0.167	0.01903	**8.78**
Itensity - AB, NoMus			
	Estimate	**Std.Error**	**t value**
(Intercept)	55.9704	1.48154	37.78
SectionB	0.66755	0.03057	**21.84**
Block	0.21033	0.01573	**13.37**
Intensity - PropChange, Mus			
	Estimate	**Std.Error**	**t value**
(Intercept)	61.52032	1.51073	40.722
PropChange	0.05114	0.06823	0.75
ConditionCont	-0.13428	0.04974	**-2.699**
Block	0.06594	0.01186	**5.561**
PropChange:ConditionCont	-0.19454	0.08311	**-2.341**
Intensity - PropChange, NoMus			
	Estimate	**Std.Error**	**t value**
(Intercept)	58.651943	1.56427	37.495
PropChange	0.047798	0.056536	0.845

ConditionCont	-0.995512	0.040218	**-24.753**
Block	-0.040185	0.009742	**-4.125**
PropChange:ConditionCont	-0.020755	0.067996	-0.305
Tempo - AB, Mus			
	Estimate	**Std.Error**	**t value**
(Intercept)	4.33386	0.27397	15.819
SectionB	-0.02705	0.08813	-0.307
Block	0.03378	0.04681	0.722
Tempo - AB, NoMus			
	Estimate	**Std.Error**	**t value**
(Intercept)	5.1765	0.21134	24.493
SectionB	-0.19957	0.07642	**-2.612**
Block	-0.04324	0.0371	-1.165
Tempo - BC, Mus			
	Estimate	**Std.Error**	**t value**
(Intercept)	4.461023	0.218536	20.413
SectionC1	-0.055892	0.096419	-0.58
SectionC2	-0.082268	0.096419	-0.853
SectionC3	0.129471	0.096419	1.343
Condition2Dur	-0.272346	0.152885	-1.781
Block	0.006523	0.028845	0.226
SectionC1:ConditionDur	0.173974	0.165547	1.051
SectionC2:ConditionDur	0.365162	0.164305	**2.222**
SectionC3:ConditionDur	0.229627	0.165547	1.387
Tempo - BC, NoMus			
	Estimate	**Std.Error**	**t value**
(Intercept)	4.938212	0.169907	29.064
SectionC1	-0.006254	0.087394	-0.072
SectionC2	-0.0116	0.087394	-0.133
SectionC3	0.076926	0.087394	0.88
Condition2Dur	-0.056632	0.12404	-0.457

Block	-0.025327	0.021399	-1.184
SectionC1:ConditionDur	-0.115835	0.147272	-0.787
SectionC2:ConditionDur	0.095072	0.146188	0.65
SectionC3:ConditionDur	-0.036382	0.147272	-0.247
Tempo - PropChange, Mus			
	Estimate	**Std.Error**	**t value**
(Intercept)	4.49046	0.21573	20.815
PropChange	-0.08149	0.11177	-0.729
Condition.Dur	-0.05561	0.12312	-0.452
Block	0.01039	0.02976	0.349
PropChange:Condition.Dur	0.10313	0.17967	0.574
Tempo - PropChange, NoMus			
	Estimate	**Std.Error**	**t value**
(Intercept)	4.97111	0.16832	29.533
PropChange	-0.04721	0.10348	-0.456
Condition.Dur	-0.1705	0.09679	-1.761
Block	-0.02151	0.02314	-0.93
PropChange:Condition.Dur	0.1936	0.16348	1.184

APPENDIX 4 (EXP.2 Production Stimuli)

we were away	a rowing rolling ram	more young wool
all year long	a yellow alley	a wall in your room
all your men	my ally	yule wool or wine
more mammalian milk	Amy Amanda	young wormy wool
a lovely yellow	arraign my ram	Yolanda warning you
a lovely lily	an eerie eel rolling	one moaning minnow
moaning minnie	my awe will, Wade	one roomy womb
my lovely yellow lemon	an enemy or Elaine	a royal Roman
your long arm	a kneeling kneeler	a rime or a roll
we are roary	an illuminary gnome	one lime yearly
I well know now	I know I knew Elenor	relying on a meme
when we were young	a lean lawn	a million oaring men

lean wrongly	a yellow lawn	an oral millionaire
nine more men	one lonely lolly	nine whirling women
lean lamb in Rome	one looming looney	a one-woman roar
more lean lamb	a lowly lamb	an oar in a winery
your lean lamb	a lonely long loom	a wooly, warm owl
our lean lamb	Lorraine or Lory	an owl in a moor
near lean lamb	a lunar lure	more men in a movie
we rule all year	lie or mail rye	a movie on a man
one male ruler	Larry or Manuel	a woman or a man
a new male ruler	mainly male looming	really, men moan more
name a male ruler	malign memory	a win in my enemy
near Manna Winery	Maury or memory	warm wine in Rome
a moaning limey	marrow rolls wrongly	all alone in Rome
we all knew	a wrong rolling memory	a one year limey
we were naming	a million mourning women	nine women in a moor
eeny meeny miny mo	women mooning a moron	one yellow or ailing ram
a lone lemming	moan away	an earring on a worm
an eery lilly	a mole or a role	a nine roar yawn
our money loan	nearly all morale	an owly yawn
minimal money on loan	long eerie wheel	a narrow loan
loan me more manatee	a rolling moon	a moaning minnow
loan me more money	I mourn lamb	a gnarly ram
we loan money	a lamb or a wall	more warring men
a narrow alley	one ram wore many	many men in a moor
a mealy worm	money or many	an eel on a moon
along a yellow wall	rain or air	a war warmed warning
all along a Roman wall	a railway	a warning in a woman
lore in a ring	rally, Norman	a warring moor
all are mine	a ray or a reign	nine men on a moon
marvel at my mama	relying on Rome	when a ram roars
a kneeling lily	roaring, rolling Ronnie	one-woman winery
a running animal	a roomy royal	nine men or women
a mainly manly roar	a rumbling rumour	in a limey whirl
a leering animal	relying on a roar	win one in a year
mainly manly women	a rumour or a rummy	win nine more
I'm mailing a mane	rolling while rooming	we are in a row

name a yellow mirror	royal rum running	row your oar
a rummy mellow	a rumour on a nail	a long narrow minnow
rhyming rolling eels	unruly wailing	who wore more wine
ailing Amanda	a wall or warm war	a winner or a whiner
arial earring	a wee whammy	many more women
a running ram	winning women	a rural ram

Appendix 5 (EXP.3 The process for extracting and scaling intensity for speech-based stimuli)

Global voice-pitch of each speaker in conversations 1b and 2b was lowered by 2 semitones, or 200 cents. This manipulation involved first extracting each speaker-channel in Praat (as described above), and then processing each as follows: A 'manipulation object' was created by selecting the sound file from the Objects window, and selecting *To Manipulation…* from the "Manipulate –" dropdown menu. The default settings of a 10ms time step, a pitch floor of 75 Hz, and a pitch ceiling of 600 Hz produced good results for all speakers. Next, a 'PitchTier' was extracted using the "Extract PitchTier" button; extracting a pitch tier object is necessary in order to manipulate voice-pitch in Praat. Selecting the newly generated 'PitchTier' item from the objects window presents new functions in the console; the "Modify –" dropdown menu contains a *Shift frequencies…* function where, after setting the appropriate time range for manipulation (i.e., the entire file), setting the 'units' to semitones and the 'frequency shift:' to -2 manipulates a PitchTier in the desired way. Highlighting the newly manipulated 'PitchTier' item along with the previous 'manipulation object' allows one to replace the PitchTier embedded within in the 'manipulation object' with the newly transformed version. Selecting the just-altered 'manipulation object' in isolation allows one to regenerate the speech using the "Get resynthesis (overlap-add)" button. This process results in a new 'Sound object' much like the original sub-recording, though with a voice-pitch lowered by the desired 200 cents/2 semitones. Each of the four channels selected for use in the pitch-altered, speech-based stimulus were altered in this way.

Once altered – or left unaltered – each group of four sub-recordings must be joined together as a single .wav file for both treatments (i.e., unaltered and pitch-lowered). Note, though, that the desired effect is babble-like in both treatments, which means all files must be played at the same time (not to be confused with concatenation, where each file would be played one after the other). To this end, the required four sound objects must all be selected simultaneously within the objects window in Praat, where the "Combine –" dropdown menu provides a *Combine to stereo* function. Because we want a mono presentation of these stimuli to more closely resemble the musical stimuli, the newly combined stereo-object is then converted into a mono sound object using the "Convert –" dropdown menu and selecting the *Convert to mono* function.

Appendix 6 (EXP.3 Additional Production Stimuli)

wally won a war	annual marina	mellow whale
wendy wore a lamb	Noel Nowhere	worrying lineman
all I name	a rainy alien	weary nominee
you're a manner	nylon-mania	warm airmail
I win, animal	manually immune	a loony rime
union role	lime warrior	honoring airmen
onion running	onion wheeler	emailing Ron
arrow rule	annoying llama	layering alimony
minor enemy	malaria	rainier lawn
lying willow	honorary owl	immemorable winger
warm owner	a lame yarn	linoleum
rural inn	whale marrow	merrily wooly
a wire mile	narrower milling	monorail
a newer nun	rhino lair	millennial awning
layer mayo	yam manor	nailing eyeliner
meanwhile leering	unaware	alarmingly oily
rely on a lion	runaway mailman	eerily whinny
airline mining	millionaire mommy	lemur nunnery