EFFECTS OF ACOUSTIC SPEECH VARIATION ON PERSONALITY TRAIT PERCEPTION

EFFECTS OF ACOUSTIC SPEECH VARIATION ON PERSONALITY TRAIT PERCEPTION

By SARA PEARSELL, MA

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2024) Hamilton, Ontario (Cognitive Science of Language)

TITLE: Effects of acoustic speech variation on personality trait perception
AUTHOR: Sara Pearsell, M.A. (York University), B.A. (University of Toronto)
SUPERVISOR: Dr. Daniel Pape
NUMBER OF PAGES: xviii, 113

Lay Abstract

This research explores the effects of different aspects of speech on the impressions of the speaker's personality. It examines three questions: (i) how loudness affects the perception of dominance, (ii) how voice quality influences personality traits, and (iii) how pronunciation variations impact charisma.

Chapter 2 (i) found that for sentences, increases in loudness increases perceptions of dominance, while for syllables they reduce them. Chapter 3 (ii) found that each voice quality investigated affects personality trait ratings, but creaky voice was perceived most negatively and smiling voice most positively. Chapter 3 (iii) found that voiced final consonants are rated higher in charisma than devoiced ones for in-person participants, but not for online participants. Regular [t] and flap pronunciations differ from glottal stops but not from each other only for online participants.

The findings suggest that certain aspects of speech variation influence personality trait ratings and offer applications to teaching and AI.

Abstract

This thesis examines acoustic properties of speech which influence perceptions of personality traits, specifically charisma. The following questions are addressed: How does amplitude variation influence ratings of dominance (i), how does voice quality affect personality trait attribution (ii), and how does allophonic variation affect ratings of charisma (iii).

Chapter 2 addresses question (i), finding that certain linguistic levels (increased amplitude in *sentence* and *syllable* levels) affected dominance ratings while others (increased amplitude at *word* level and reduction at *syllable* level) did not. Increased *sentence* amplitude increased dominance ratings while increased *syllable* amplitudes had inverse effects. Additionally, two types of dominance were examined (*social* and *physical* dominance) but no statistically significant differences were found between the two.

Chapter 3 examines question (ii). All voice qualities investigated (modal, creaky, breathy, nasal, and smiling) were found to be statistically significant. Effect sizes for statistical significance varied for each voice quality. Creaky voice (rated the lowest/ most negative) and smiling voice (rated the highest/most positive) had the strongest effects.

Chapter 4 examines question (iii). Experiment 1 (in-person) and Experiment 2 (online) examined the effects of allophonic variation, final consonant devoicing (FCD), and /t/ variation, on ratings of charisma. Experiment 1 found statistically significant rating differences for FCD. Final voiced items were rated higher compared to devoiced ones. For the /t/ variation, only speaker differences were found to be statistically significant. Experiment 2 showed no statistically significant results for FCD, whereas /t/ variation found statistical significance for [t] productions versus the glottal stop, and for flap productions versus the glottal stop. No rating differences were found between [t] and flap.

Overall, this thesis demonstrates that some acoustic variations within speech affect personality trait ratings, specifically charisma, while others do not. I discuss reasons for these outcomes and their utilization in various domains, including AI.

Acknowledgements

This academic endeavour has felt arduous and at times never-ending. Without the support, guidance, and thoughtful advice of so many I would not have been able to complete this huge accomplishment and because of this support, I would like to thank everyone who was a part of this thesis.

Firstly, I would like to thank my supervisor, Daniel Pape. Daniel, I know our academic beginnings were less than ideal at times, but I am so grateful you never gave up on me even when I felt completely discouraged. Your vigour for academia, your acute eye for detail, and your passion for furthering the education of graduate and undergraduate students alike is more than inspiring. Your fervor for academia has played a pivotal role in my present accomplishment of completing this thesis. I truly would not be here without you. The skills I have acquired during this Ph.D. are thanks to you and I hope to continue to grow in these skills as I continue my academic journey. Thank you for believing in me and I hope, despite the rough start we had, I have made you proud!

I would also like to thank my committee members, Magda Stroinska, and Elisabet (Lili) Service. Their support alongside their feedback, suggestions, and new perspectives on the research within my thesis and in my other academic pursuits have expanded my own perspective both in and outside the field of academia. Without their support, encouragement and insight, I would not have been able to accomplish this undertaking.

I would also like to thank all the other professors who have inspired new skills, and insightful viewpoints, whether from lecturing in the classroom, discussions on current research, or provision of insightful commentaries.

Thank you to all my fellow Phonetic Lab members who always encouraged me and made me laugh when I was overwhelmed. I would like to specifically thank Carla Weigel, and Monika Krizic who have been amazing colleagues turned friends, offering me company and support many times over during breakthroughs and breakdowns. Without their kind words and support, I would not have been able to complete this dissertation.

Lastly, I would like to thank my late cat, Lemon. She came to me shortly after the passing of my father and before my solo move to Toronto where I pursued my undergraduate education. She provided countless hours of comfort, companionship, friendship, and love and pushed me through some of my hardest days, providing me with the drive to never give up my academic dreams. Although she did not get to be a part of the completion of *this* milestone, I will never forget all the other milestones she was a part of, like my master's degree, all the conferences, presentations, and countless all-nighters. I know that she sent my two other cats into my life, Juno and Nickel, and will forever be appreciative of their presence in my life.

To those who dare to fail miserably,
may you always achieve greatly.

Table of Contents

References 112

List of Tables

List of Figures

CHAPTER 2

CHAPTER 3

CHAPTER 4

Declaration of academic achievement

This is a "sandwich" thesis, as defined by the School of Graduate Studies at McMaster University. It contains three empirical studies of which I am the primary author. In the following sections are the details and roles of each of the authors for the three studies within this dissertation.

Chapter 2

This study has been submitted and is currently in-print in *LACUS Forum, volume 48* as **Pearsell, S.**, Pape, D. The effects of different levels of amplitude variation on the perceived dominance of a speaker.

Pearsell, S.: study design, stimuli recording, stimuli preparation, literature review, data collection, data analysis, manuscript writing, revision, preparation for publication, peer review revisions.

Pape, D: study design, stimuli preparation, data analysis, manuscript writing and revision, peer review revisions.

Chapter 3

This study has been submitted and published in *Frontier's in Communication: Psychology of Language* as **Pearsell, S.**, Pape, D. (2023). The effects of different voice qualities on the perceived personality of a speaker.

Pearsell, S.: study design, stimuli recording, stimuli preparation, literature review, data collection, data analysis, manuscript writing, revision, preparation for publication, peer review revisions.

Pape, D: study design, stimuli preparation, data analysis, manuscript writing and revision, peer review revisions.

Chapter 4

This study has been prepared for publication but has not currently been submitted as **Pearsell, S.**, Pape, D., Service, E. (2023). The influence of allophonic variation on the perceived charisma of a speaker.

Pearsell, S.: study design, stimuli recording, stimuli preparation, literature review, data collection, data analysis, manuscript writing, revision, preparation for publication.

Pape, D: study design, stimuli preparation, data analysis, manuscript writing and revision.

Service, E.: data analysis, manuscript writing and revision.

Additional achievements

In addition to the studies presented in this dissertation, the author contributed to the following conference proceedings:

Pearsell, S. & Pape, D. (2023, Aug 20). *The Effects of Different Levels of Amplitude Variation on Perceived Speaker Dominance*. In: Radek Skarnitzl & Jan Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences (pp. 206–210). Guarant International.

CHAPTER 1

Introduction

1.1 Aims of the Current Thesis

Speech is one of the most crucial facets of human communication. At its core, speech production involves the coordinated activation of various anatomical structures, from the larynx to the lips, all coordinated by the intricate neural networks within the speech processing centers of the brain. Outside of comprehension, the acoustic signals of speech play complex roles in how humans exchange information, communicate emotions, and intentions, essentially forming the cornerstone of interpersonal communication. Apart from making comprehension possible, the acoustic speech signal always carries information about specific and consistent differences, such as consistent amplitude differences or voice quality differences.

Some of the variations present in speech will be treated as noise by the listener while other consistent variations across speakers or even within speakers can manifest in robust perceptual judgment differences. These perceptual types of variations can be observed in many ways, both across speakers, for example the vowel variations between British English speakers versus Canadian English speakers, as well as within speakers, such as the voice quality produced by a singular speaker in an important business meeting versus the voice quality produced when with a group of friends. This leads to questions regarding these variations: are these consistent differences used to generate impressions about that speaker's personality, or are they just treated as variability that does not contribute to the perceptual domain? Previous research suggests the former. This will be reviewed in forthcoming chapters.

The mechanisms of speech manipulation and its utilization to convey characteristics about our personality or attributes have seemingly always been of interest to both popular culture and academic research. Pittam (1985) explored the effect of voice quality on listener perceptions of solidarity, attractiveness, and speaker status. Voice quality differences emanate from the combination of *laryngeal* and *supralaryngeal* features continuously present in a speaker's speech production, resulting in the specific auditory features, such as tone, of a speaker. Pittam found several relationships between different voice qualities and their elicitation of rating scores across different personality traits. For example, the author found that the presence of breathiness, or whispery qualities in a speaker's voice increased the ratings of solidarity for listeners, whilst perceptions of tense voice and breathy voice increased perceptions of speaker status. Bosker (2021) investigated the effects of amplitude modulations linked to syllabic speech rate and found that more pronounced amplitude modulations increased charisma ratings. Lower fundamental frequency (f0) has been found to increase perceptions of dominance in speakers (Puts et al., 2007).

The above-mentioned studies are only a small sample of the previous research which has explored various facets of acoustic variations (differences) and their effects on the perception of traits associated with a speaker. Like the copious number of acoustic features available for analysis, such as those presented in the previously mentioned research (voice quality, amplitude, f0), there seems to be an unexhaustive collection of personality traits which can also be explored, such as persuasiveness, intelligence, charisma, etc. This creates a twofold problem: which acoustic differences have salient and observable effects and how can personality traits be categorized in meaningful ways to capture these potential acoustic effects?

The first problem, pertaining to which acoustic differences are observable and which of these are treated as noise is a rather large one. As mentioned in the previous literature, several aspects appear to produce information meaningful to listeners with respect to speaker personality trait ascription. To narrow down the scope of the current research, three acoustic speech aspects have been selected to be varied and analyzed in the following chapters of this thesis. The first speech aspect is variation in amplitude (loudness) (Chapter 2), the second is differences in voice quality (Chapter 3), and the third is allophonic variation (Chapter 4). These three distinctive speech dimensions capture a spectrum of production locations in the vocal tract (i.e., voice quality: supralaryngeal & laryngeal; amplitude: laryngeal; and allophony: supralaryngeal). These three acoustic dimensions also allow examination of the effects of suprasegmental differences (i.e., amplitude, and voice quality) as well as inter-speaker versus geographical/sociophonetic differences (i.e., allophonic variation). The social effects of each of these aspects have practical applications in real world settings, such as office or work presentations, while also being relevant for clinical application in speech language pathology and speech therapy.

The second problem originates from the unexhaustive list of personality traits. Frequently, personality traits are vaguely categorized or defined, leaving open the question of what researchers are actually examining. This is particularly relevant to the personality trait of *charisma.* Definitions of charisma often lack the empirical explanatory power to clarify what charisma or charismatic traits actually mean. In order to address this question, this thesis attempts to measure and define charisma in a more empirical manner than previously explored.

This thesis focuses on how manipulation of specific speech characteristics affects perception of a speaker's personality traits that contribute to the perception of *charisma*. More specifically, the following chapters focus on how voice quality, voice amplitude, and allophonic variation affect the perception of charisma. These three specific variables offer insight into the effects of laryngeal acoustic differences (i.e., amplitude), supralaryngeal differences (i.e., allophonic variation), as well as the combination of both laryngeal and supralaryngeal differences (i.e., voice quality) on speaker perception. In part, this provides a diverse exploration of acoustic signals in terms of the vocal tract location. Moreover, these speech characteristics of interest also offer insight into the effects of segmental and geographical/sociophonetic differences (i.e., allophonic variation) and suprasegmental (amplitude, voice quality) domains.

One major gap in the previous literature is the comparison of acoustic differences across speakers as well as within speakers. Previous studies (Laver, 1980; Pittam, 1985; Puts et al., 2007; Rosenberg and Hirschberg, 2009; Anderson et al., 2013) have often only investigated across-speaker, but not within-speaker variation. Situations and environments impact *how* we produce speech. Through the introduction of novel dimensions, such as new factors of interest, as well as different methodological approaches, the following experiments expand upon previous research as well as fill current gaps in the literature.

The issues covered in this thesis are summarized by the research questions below.

1. How do different levels of speech amplitude variation affect the perceived dominance of a speaker?
2. How does voice quality affect listener ratings of the personality of a speaker?
3. How does allophonic variation influence the perceived charisma of a speaker?

Exploration into these topics will contribute to phonetic and speech research on personality trait ascription to speakers. We varied voice quality productions within a speaker, amplitude variation across different linguistic levels (i.e., sentence, word, and syllable), and allophonic variation (in particular, final consonant devoicing and /t/ phoneme variation). In the following sections, we introduce some basic concepts and explain how they will contribute to answering the research questions posed in this thesis.

1.2 Specific Research Questions of the Present Thesis

1.2.1 How do different levels of speech amplitude variation affect the perceived dominance of a speaker?

Chapter 2 examines the first acoustic area of interest within this thesis: the effects of speech amplitude variation on perception of dominance in speakers. Previous research has shown that increases in amplitude (Scherer et al., 1973; Harrigan et al., 1989), loudness (Buller & Burgoon, 1986), and intensity (Aronovitch, 1976) boost ratings of dominance in a speaker. Tusing & Dillard (2000) found that increase in overall mean vocal amplitude in spontaneous speech positively correlates with ratings of speaker dominance. More confident individuals have also been found to produce speech with greater amplitudes (Kimble & Seidel, 1991). Furthermore, high signal intensity is linked to perceptions of dominance (Aronovitch, 1976). Despite these studies, the research on this topic is particularly limited. One crucial component missing from these studies is the impact of linguistic level, from syllables to phrases, in terms of amplitude manipulation. Another pivotal limitation of all these studies lies in the lack of conceptual clarity regarding the multifaceted contrast of *dominance.*

Dominance is commonly defined as some variation of influence or power over others. This definition is exceptionally vague as dominance could relate to features of influence and power in social ability (such as charisma in a speaker) or physical attributes. To address this, the current study opts to follow definitions offered by Mueller and Mazur (1997, p. 570). They provide the following description which encapsulates their definition of *social dominance*: “a dominant person tells other people what to do, is respected, influential, and often a leader, while submissive or subordinate people are not influential or assertive and are usually directed by others”. The authors’ definition of *physical dominance* is determined by ratings of speaker's ability to win a physical fight. By these definitions, *social dominance* involves characteristics that are more socially based and could be linked to personality traits such as *charisma* – the focus of interest for personality traits in this thesis – while *physical dominance* is self-defining in its association with physical power or influence. Through the differentiation between dominance types, the present study starts to form a deeper understanding of *charisma* characteristics and whether comparisons between these two types of dominance provide deeper insight into what it means to have charisma.

The study described in Chapter 2 builds on the previous research of Puts et al. (2007) with a focus on speech amplitude differences and their effect on dominance ratings. Amplitude (the magnitude of a sound waves vibration which is directly related to the loudness or volume of the sound) and intensity (the power carried by a sound wave per unit area) are acoustic parameters of sound while loudness is a related perceptual parameter. Thus, in Chapter 2, the current study examines the acoustic parameter of amplitude and the influence of its variation on three different levels of language: (i) *sentence/paragraph* (i.e., amplitude enhancement of the entire sentence or paragraph), (ii) *word* (one specific word is put in focus, realized with enhanced amplitude), and (iii) *syllable* (i.e. increasing/decreasing amplitude differences between all stressed versus unstressed syllables). The investigation of three levels moves across different linguistics domains from high-level linguistic domain (i.e., sentence) to low-level linguistic domain (i.e., syllable) and aims to find which of these linguistic levels are the most salient in higher listener ratings of dominance (i.e., higher ratings equal more positive perceptions, lower ratings equal more negative perceptions). The results for the low-level syllable domain in particular provide insights into an unstudied linguistics level. Supplementarily, the examination of all three domains (sentence, word, syllable) simultaneously provides novel insights into the effects of acoustic amplitude differences on social dominance ratings. As social dominance appears to correlate with features typically associated with charisma, the results of social dominance can thus attempt to further the understanding of charisma perception in speakers.

The results from the study described in Chapter 2 showed that increases in amplitude for the *sentence* condition positively affected perceptions of both social and physical dominance. In the *syllable* condition, enhances in amplitude had the inverse effect, decreasing ratings of dominance for both social and physical dominance. No significant effect was found for the reduced *syllable* condition or the *word* condition. Investigation into the different linguistic levels reveals a more detailed image of how amplitude influences the perception of dominance.

Firstly, the results demonstrate that not all amplitude increases lead to higher ratings of dominance. Ratings depend on the linguistic level manipulated. This suggests that amplitude variation at some linguistic levels is treated as noise, as seen in the reduced syllable condition or enhanced word condition Others (i.e., enhanced *syllable* condition and enhanced *sentence* condition) carried important perceptual information relevant to listeners and their determination of personality traits within speakers. Another key result is the lack of differentiation between *social* and *physical* dominance in terms of perceptual ratings. This in an interesting result, suggesting that dominance understood as a more general concept is sufficient for perceptual dominance ratings. In turn, this also indicates a potential correlation between dominance and charisma traits.

1.2.2. How does voice quality affect listener ratings of personality for a speaker?

The next question of interest within this thesis relates to the effects of voice quality on ratings of speaker personality. Chapter 3 builds on the previous research regarding voice quality. Prior studies investigated the relationship between voice quality within speech productions and its connection to personality trait perception. As mentioned above, Pittam (1985) examined the effects of different vocal qualities on listeners' ratings of solidarity, attractiveness, and status of the speaker. The results from their study found that listeners' ratings of solidarity with a speaker were greater when there was the presence of either *breathiness* or *whispery* voice. Additionally, the study found that ratings of speaker status were higher for *tense* voices, as well as voices with *breathy* quality, compared to voices which had *whispery* or *nasal* qualities. A study by Anderson et al. (2013) demonstrated that *creaky* voice in women negatively impacted ratings of competence, trustworthiness, and education level. Laver (1980) also found a link between perceived sexuality and sensuality in female speakers when the voice had a *breathy* quality. This effect was not found in male speakers. Although these studies, among others in the field of voice quality research, provide a starting point of reference for the present study, one larger and important component is missing from this research: within-speaker (i.e., intra-speaker) effects.

Voice quality can be affected by several situational factors such as physiological afflictions, e.g., a pathology causing the production of a creakier or harsher voice quality or environmental impacts, e.g., opting for a more whispery voice (in females) to sound more seductive. Much like the finer grain examination of the various linguistic domains and amplitude variants in Chapter 2, Chapter 3 aims to investigate the complex intricacies of within-speaker and across-speaker effects of variability in speech. Currently, there remains a gap in the research regarding within speaker effects of voice quality differences with previous research focusing primarily on across-speaker effects. Therefore, the question of whether within-speaker voice quality changes affect speaker perception remains to be answered. One of the objectives of this study is to attempt to shed light to this topic.

Traditionally, voice quality has been defined as "the quasi-permanent quality of a speaker's voice" (Abercrombie, 1967) and "those characteristics which are present more or less all the time that a person is talking. It is a quasi-permanent quality running through all the sound that issues from his mouth". Later research corroborates this definition, expanding that differences between voice qualities are the result of both laryngeal and supralaryngeal features which produce specific auditory colouring in an individual's voice (Laver, 1980; Esling et al., 2019). Conceptually, voice quality falls on a continuum with breathy voice (produced with a more open glottis) on one extreme of the continuum, and creaky voice (produced with a more constricted glottis) on the other end of the scale. Modal voice (produced with a moderate constriction of the glottis) falls in between the two. Modal voice has been described as the "neutral mode of phonation"

(Laver, 1980, p. 110). This traditional description of voice quality focuses primarily on the laryngeal features in speech production. Although we also adopt this definition, we have opted for the term *vocal quality* in our study to stress that this term is inclusionary of both laryngeal features found in traditional voice quality descriptions (e.g., modal, creaky, and breathy voices) but also continuous suprasegmental features such as smiling and nasality, two other dimensions of interest we investigate in Chapter 3.

The previous research on voice quality has investigated individual voice qualities in a speaker's voice, examining one voice quality (e.g., exclusively creaky voice) at a time in one speaker. This is problematic when viewing the larger scale implications of speaker variability when it comes to speech production. As humans, our voice quality varies dependent on the circumstances or environments in which we are speaking. By examining only one vocal quality per speaker, information regarding the saliency of each voice quality comparative to other voice qualities, as well as effect sizes remain unaddressed. This gap in the present literature sets the premise for the experiment conducted in Chapter 3 of this thesis, which examines the production of several aspects of vocal quality within a speaker while simultaneously investigating multiple speakers.

Another novel aspect of the study in Chapter 3 is the examination of smiling and nasal voice (specifically hypernasality) in combination with modal voice productions. Specifically, we also investigated if there are any differences such as a ceiling effect between *normal* versus *extreme smiling* conditions as found in Tschinse et al. (2022), as well as any difference between *natural* versus *technical* breathy voice. The *technical* breathy voice is created by taking the modal voice production and overlaying a speech-shaped noise signal onto the stimuli. This results in a new synthesis file with harmonics-to-noise ratios (HNR) corresponding to natural breathiness in voices but distributed in a much more technical and non-realistic manner onto the acoustic signal, thus presenting an interesting case-study comparing perception of a natural breathy voice with one created by addition of technical noise.

Listening to stimuli capturing vocal qualities of interest (modal, creaky, breathy, nasal, and smiling voice), participants rated them on continuous sliding scales linked to personality traits associated with *charisma*. Although the characterization of *charisma* as a personality trait remains speculative (Michalsky & Niebuhr, 2019), studies have demonstrated a relationship between charismatic features within speech and the *Big 5 Personality Traits* (Antonakis et al., 2016; Michalsky & Niebuhr, 2019). Therefore, we modelled and modified the scales in this study from previous research which used *The Big 5* as the premise for their own research (Puts et al., 2007; Rosenberg and Hirschberg, 2009). The *Big 5 Personality Traits* (or the *Big 5 of Personality*) is a prominent psychological theory of personality traits (Norman, 1963; McCrae and John, 1992) which categorizes personality traits into five dimensions: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (or OCEAN for short). Each dimension represents a range from high scores on one side (e.g., the subject demonstrates extroversion) and low scores on the other (e.g., the subject demonstrates introversion). Based on the previously mentioned research by Puts et al. (2007) and Rosenberg & Hirschberg (2009), we opted to exclude the trait of *openness* from our study due to a lack of established relationship to ratings of *charisma.*

This leads to the secondary objective of the study presented in Chapter 3: to propose a more concrete and empirical definition of *charisma*. As is discussed in Chapter 2, definitions of charisma remain vague and are not formulated in a conceptually explicit manner. Chapter 2 examines two different types of dominance, *social* and *physical*, with the working hypothesis that *charisma* falls under the scope of *social* dominance. However, the results from the study in Chapter 2 found that there were no differences between *social* and *physical* dominance in terms of ratings. Although the results provided interesting insights into listeners' perceptions pertaining to concepts of dominance in a more general regard, they did not provide any further insight into *charisma* and charismatic traits. To achieve a less vague and more empirical approach to classification of *charisma*, Chapter 3 focuses on methods utilized in previous studies (Puts et al., 2007; Rosenberg & Hirschberg, 2009) combined with the previously well-established psychometric scales for personality (i.e., *The Big 5 of Personality)*. Through the combination of these two approaches, we aimed to empirically categorize a method of measuring charisma, one based on previously established paradigms. This approach is novel in the field of personality traits, and, more specifically, charisma, research. By adopting this approach, quantification of charisma using the *Big 5* allows for more targeted understanding regarding which attributes form various trait categories to create the concept of charisma. This, in turn, may assist in a more general interpretation of personality trait perception using the *Big 5*.

Taking all the above variables into consideration we examined voice quality variation in two ways. First, we compared within-speaker differences. To analyze these results, we compared the differences of each produced voice quality to the "baseline" (i.e., *modal* voice) for each individual speaker. This allowed for more accurate representations for each speaker as the comparisons were made against their own voice rather than to other voices where other features may impact the personality ratings of listeners. For example, the f0 or general timbre of a speaker could override voice qualities of interest. By comparing the voice quality of a speaker against their own modal voice, external acoustic factors which vary across speakers were largely controlled.

The experiment in Chapter 3 found that listeners rated creaky voice significantly lower for all personality traits. This was found as a trend within, and across all speakers. Smiling voice produced the opposite effect. Smiling had significantly more positive (higher) ratings for all personality traits. These results suggest that general increases or decreases in ratings for each individual personality trait from the *Big 5* across vocal qualities were linked to perceptions of charisma. As the results demonstrated similar trends across all personality traits each personality trait of the *Big 5* can be seen singularly reflective of ratings for charisma. Furthermore, creaky voice negatively influenced perceptions of charisma in speakers, while smiling positively influenced charisma ratings. Nasal voice and breathy voice also demonstrated statistically significant results from modal voice, however, their overall impact on ratings was less pronounced than those of creaky voice and smiling. The results also identified gender trends. Female speakers received more negative ratings for creaky voice but higher ratings for smiling when compared to their male counterparts.

Some novel results from this study are the demonstration that certain vocal quality types have more saliency than others among all statistically significant results. For example, although nasal voice negatively impacted listener perceptions of personality traits, creaky voice had a stronger negative effect.

1.2.3. How does allophonic variation influence the perceived charisma of a speaker?

The final area of acoustic variation of interest for this thesis pertains to allophonic variation and its influence on the perceived charisma of a speaker. Chapter 4 examines allophonic variation in two ways. The first variation phenomenon is final consonant devoicing (FCD). The second variation type is the allophonic alternation of the voiceless alveolar plosive /t/ with flapping in specific environments, particularly in the North American phonological environments. Currently, no research specifically explores the role allophone production plays in personality trait attribution. Moreover, there is no research which examines its role in charisma ratings. The study presented in Chapter 4 provides exciting, novel research in the exploration of acoustic differences and their impact on perception. This chapter aims to reveal insights into another set of acoustic features which may have a role in personality trait attribution.

Chapter 4 consists of two experiments which build on each other. The initial experiment examined personality ratings by in-person participants, whilst the latter experiment focused on online participants. With the increase in usage of online communication platforms in daily life, from work meetings to online lectures, understanding the acoustic implications of speech perception in an online environment becomes ever more relevant. Exploration of both presentation formats can provide a deeper understanding of the psychological impact of speech related variables as well as the acoustic effects of technology integration on the speech signal.

An allophone is labelled as one of potentially several speech sounds (a phone) which is used to pronounce a singular phoneme. For example, the phoneme /p/ in English can manifest with the two alternative phones: the unaspirated [p] and aspirated [p^h]. The alternation of allophonic variants does not change the meaning of a word. Allophones often exist in complimentary distribution – in contexts where one variant is realized, another variant of the same phoneme cannot. Essentially, the phonotactic environment of an utterance dictates which variant is produced, such as [p] when the phoneme /p/ is not word initial as in the word *spin* [spɪn] or [p^h] when it is in initial position, such as in the word *pin* [p^hɪn]. Other factors outside phonotactic environment can also influence which allophonic variant is selected and produced. Examples of these other factors may include geographical or dialectal regions of a particular language (e.g., English from Toronto (Ontario) versus English from London (England)), the sociolinguistic background of a speaker (e.g., the English used by higher socioeconomic classes versus that used by lower socioeconomic ones), and language background (e.g., a Spanish non-native English speaker versus a native English speaker). As a result of some of these factors, the production of an allophone or allophones may result in uncommon or non-standard varieties of spoken language.

Due to the complexity of allophones and their distribution, both across and within languages, questions regarding their sociolinguistic function arise. One of these questions and the focus of Chapter 4 is the how allophonic alternation may impact perceptions of charisma. This chapter specifically focuses on two allophonic environments. The first environment, as previously mentioned, is final consonant devoicing (FCD). FCD is the phonological process in which the voiced final consonant (e.g., /v b d g/) of a word becomes voiceless (e.g., /f p t k /). This process changes the production of a word such as 'had' (/hæd/) to sound like 'hat' (/hæt/). Although this process is not present in standard varieties of North American English, it can be found in some non-standard English dialects such as African American Vernacular English (AAVE). Additionally, it can also be found in the speech of non-native English speakers whose native language contains this phonological process (e.g., German, or Russian).

The second allophonic variation of interest is /t/ variation, specifically allophones of the voiceless alveolar stop/plosive (/t/). In Canadian English, there are several phonological variants of the underlying phoneme /t/ in certain environments. For the study in Chapter 4, the environments of interest for the /t/ allophones of interest are generally traditional "flapping" contexts. For this study, the environments investigated are summarized as: i) between two vowels (including syllabic sonorants; Walker, 2015; e.g., 'city' /sɪ**ɾ**i/), ii) following a nasal and preceding an unstressed vowel (e.g. 'wanted' /wɑn**ɾ**əd/, and iii) across word and syllabic boundaries when /t/ is in the coda position of the first word/syllable and an unstressed vowel is in the onset position of the second word/syllable (e.g., 'my cat is grey' /maɪ kæ**ɾ** ɪz gɹeɪ/. Essentially, these environments restrict /t/ from occurring in a stressed position in order for the allophone variants of the phoneme [t] to be produced.

As previously noted, there are currently no studies which examine personality trait perception for variation in allophone production. One study by Niebuhr (2017) does suggest speech reductions in prosodic and segmental domains affect charisma ratings with speech reductions. These results demonstrate systematic effects as a consequence of speech utterance reductions. Although not directly related to allophones, the evidence of systematic outcomes in Niebuhr's study, at least regarding the segmental level, does suggest there may be similar outcomes as result of the nature of allophones. As allophones function in complementary distribution on a segmental level, it could be inferred that they would also behave systematically and thus produce similar results to the study by Niebuhr. Therefore, the results found by Niebuhr provide some valuable insight into potential outcomes for the study in Chapter 4.

Niebuhr investigated several aspects of speech reduction involving segmental and prosodic components. These included naturally produced systematic variation within sentences. The stimuli consisted of a spectrum of (1) fully produced sentences, (2) slightly reduced productions (e.g., informal conversation), and (3) strongly reduced productions (i.e., content and function words were greatly reduced in production). Focusing on the segmental results, the study found that the stronger the reduction in the sentence, the greater the probability of listener-perceived speaker attributions of absent-mindedness or clumsiness. Furthermore, speakers were also observed as less skilled, less sociable, less optimistic, and less educated. Although Niebuhr's study was not directly correlated to perceptions of allophonic variation, it does show the impact systematic phonemic or prosodic variation may have on perceptions of charisma. These results provide groundwork evidence that systematic variation in speech output, as found in segmental reduction, impacts listener ratings of speaker personality. With this knowledge, it is expected that similar results could be found for allophonic variations, hence the motivation for the current study.

Chapter 4 contains two separate experiments using the same stimuli from six different speakers; one experiment was conducted *in-person* and the other one was conducted *online*. Each experiment was subdivided into two components: the effect of final consonant devoicing and /t/ variation on personality ratings, respectively. For the first experiment, listeners (participants) were recruited using the SONA software at McMaster University. They completed the experiment in-person on an iMac computer using high quality headphones and audio interface in a sound treated room. As a result of these conditions, there were minimal if any environmental distractions and the environment was optimal for auditory input. For the online experiment, participants were recruited through and online platform Prolific and used their own computer and whichever headphones or earbuds were available to them. Although advised to eliminate distractions and focus solely on the experiment, complete focus and suitable acoustic conditions were, thus, not guaranteed.

Individually, these studies offer new insights into the effects of allophonic variation. For the case of FCD this is new perspectives on the effects of non-native phonological patterns and on charisma ratings and personality ratings more generally. For /t/ variation, these perspectives can provide insight into the effects of regional/sociophonetic variation on charisma scores.

1.3 Outline of the Thesis

Chapter 2 reports the effects of amplitude manipulation at different linguistic levels on the perception of speaker dominance. That chapter compares three different linguistic levels (syllable, word, and sentence) on two different classifications of dominance (physical and social) with social dominance being representative of charismatic personality traits. Chapter 3 reports the effects of voice quality on the perception of speakers' personality traits. Within the chapter, the effects of seven different voice qualities are compared on the ratings of different personality traits in the *Big 5 of Personality* and correlated to definitions of *charisma.* Chapter 4 reports the effects of allophonic variation on the perceptions of charisma. This chapter looks at two variants of interest: final consonant devoicing and /t/ variation. Additionally, the chapter is divided into two experiments with the first examining these allophonic variants for *in-person* participants and the second examining *online* participants. Lastly, Chapter 5 summarizes and discusses the research findings and suggests lines of further research in this field.

References

Abercrombie, D., 1967. Elements of General Phonetics. *Edinburgh: Edinburgh University Press.*

Anderson, R.C., Klofstad, C.A., Mayew, W.J. and Venkatachalam, M., 2014. Vocal fry may undermine the success of young women in the labor market. *PloS one*, *9*(5), p.e97506.

Antonakis, J., Bastardoz, N. & Jacquart, P. 2016. Charisma: an ill-defined and ill-measured gift. *Annual Review of Organizational Psychology and Organization Behaviour,* 3, pp.293-319.

Aronovitch, C.D., 1976. The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *The Journal of social psychology*, *99*(2), pp.207-220.

Bosker, H.R., 2021. The contribution of amplitude modulations in speech to perceived charisma. *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*, pp.165-181.

Buller, D.B. and Burgoon, J.K., 1986. The effects of vocalics and nonverbal sensitivity on compliance: A replication and extension. *Human Communication Research*, *13*(1), pp.126-144.

Esling, J.H., Moisik, S.R., Benner, A. and Crevier-Buchman, L., 2019. *Voice quality: The laryngeal articulator model* (Vol. 162). Cambridge University Press.

Harrigan, J.A., Gramata, J.F., Lucic, K.S. and Margolis, C., 1989. It's how you say it: Physicians' vocal behavior. *Social Science & Medicine*, *28*(1), pp.87-92.

Kimble, C.E. and Seidel, S.D., 1991. Vocal signs of confidence. *Journal of Nonverbal Behavior*, *15*, pp.99-105.

Laver, J., 1980. The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, *31*, pp.1-186.

McCrae, R.R. and John, O.P., 1992. An introduction to the five-factor model and its applications. *Journal of personality*, *60*(2), pp.175-215.

Michalsky, J., & Niebuhr, O. 2019. Myth busted? Challenging what we think we know about charismatic speech. *AUC PHILOLOGICA, 2019(2),* pp.27-56.

Niebuhr, O., 2017. Clear Speech-Mere Speech? How segmental and prosodic speech reduction shape the impression that speakers create on listeners. In *Proceedings of the International Conference on Spoken Language Processing* (Vol. 18, pp. 894-898). International Speech Communication Association.

Norman, W.T., 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The journal of abnormal and social psychology*, *66*(6), pp.574.

Pittam, J., 1985. Voice quality: its measurement and functional classification.

Puts, D.A., Hodges, C.R., Cárdenas, R.A. and Gaulin, S.J., 2007. Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), pp.340-344.

Rosenberg, A. and Hirschberg, J., 2009. Charisma perception from text and speech. *Speech Communication*, *51*(7), pp.640-655.

Scherer, K.R., London, H. and Wolf, J.J., 1973. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, *7*(1), pp.31-44.

Tschinse, L. M., Asadi, A., Gutnyk, A., & Niebuhr, O. 2022. Keep on smiling...? On the sex-specific connections between smiling duration and perceived speaker attributes in business pitches. In Proc. 11th International Conference of Speech Prosody, pp.1-5.

Tusing, K.J. and Dillard, J.P., 2000. The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence. *Human Communication Research*, *26*(1), pp.148-171.

CHAPTER 2

THE EFFECTS OF DIFFERENT LEVELS OF AMPLITUDE VARIATION ON THE PERCEIVED DOMINANCE OF A SPEAKER

This study has been submitted and is currently in-print in *LACUS Forum, volume 48* as Pearsell, S., Pape, D. The effects of different levels of amplitude variation on the perceived dominance of a speaker.

Abstract

This paper examines how varying suprasegmental amplitude of speakers' productions on different linguistic levels affects listeners' dominance ratings of that specific speaker. 6 English speakers produced audio stimuli to be judged in a perception experiment. The stimuli's amplitudes were manipulated on three different levels: (i) *sentence* (entire phrase/paragraph), (ii) *word* (one specific word in focus), and (iii) *syllable* (stressed/unstressed syllables). Participants rated 4 statements on sliding scales examining *physical* and *social* dominance. Results showed *sentence* level manipulations received the highest dominance ratings with significantly higher ratings compared to the baseline condition. Contrastively, *word* and *syllable* level intensity changes generated lower dominance ratings (compared to baseline), however only the *syllable* level manipulation resulted in significant rating differences. Examinations of the effect of *speaker* and *listener* gender showed that only *physical* dominance ratings on the *syllable* level was significant comparing female and male *listeners*.

Keywords: amplitude variation, perceived personality traits, speech perception, variability in speech

IN RECENT RESEARCH (Quené et al., 2016; Berger et al., 2017; Yang et al., 2020; Rafikova et al., 2022), there has been an increase of interest in examining the acoustic properties of speakers' voices and the connection these acoustic properties have with perceived personality trait attribution (of that same speaker). Previous research has observed the influences of pitch (Puts et al., 2007; Rosenberg & Hirschberg, 2009; Quené et al, 2016; Berger et al., 2017), pause-filled-gaps (Möbius, 2003; Niebuhr & Fischer, 2019), etc., on the perceived vocal and/or personality traits of a speaker such as charisma, dominance. However, to our knowledge, very little work has been done on the effects of varying suprasegmental amplitude differences, or reversely their perceived loudness of these parameters, on these traits[1].

1. PREVIOUS RESEARCH. With respect to amplitude and intensity differences in the speech signal previous research has demonstrated how (1) amplitude (Harrigan et al., 1989; Scherer et al., 1973), (2) loudness (Buller & Burgoon, 1986), and (3) intensity (Aronovitch, 1976) are positively correlated to perceptions of dominance. Scherer (1974) examined the effects of amplitude on potency judgements. These potency judgements are a construct which shows overlap conceptually with the definitions of dominance. Research by Tusing & Dillard (2000) found that mean vocal amplitude is positively associated with dominance ratings for the production of spontaneous speech: the louder the produced vocal amplitude, the higher the perceived dominance ratings. In their study the speech production was a short audio message and examined the amplitude (and f0), over several phrases. This experimental design essentially examined the amplitude of the whole phrase or sentence. In their study, dominance was classified more generally as "When one person dominates another, that person has control or power over the other" and did not make a distinction between difference dominance types (e.g. social or physical dominance). Additionally, the researchers only manipulated whole audio messages which would correlate to a whole phrase environment.

Other research has found that more confident individuals speak with greater amplitude (Kimble & Seidel, 1991). In this study, participants were asked a series of questions and verbally responded their choice from multiple answers. Depending on which block participants were in, there would be 0 to 2 experimenters present in the room during their answers. Participants then had to score how confident they were in their answers. The recorded participant audio was converted to amplitude measurements (integrated, amplified, and digitized to output values for amplitude with silent periods not affecting these values) to determine the loudness of these productions. The values from the average loudness were linearly correlated to signal amplitudes, not logarithmically correlated like decibel values (intensity). The research then compared the amplitude measurements of these verbal responses to the "confidence" scores participants gave themselves. Since these verbal answers were short, one or two words, it could be argued that this study was looking at word stress. Additionally, because the researchers were

[1] Please note that amplitude and intensity differences are acoustic parameters, whereas loudness is a purely perceptual measure, but of course tightly linked to amplitude/intensity differences.

correlating the confidence of these answers, this research would be classified into a social dominance category.

Lastly, high signal intensity is associated with perceptions of dominance (Aronovitch 1976; Scherer et al. 1973). In the study by Aronovitch (1976), participants listened to short audio recordings (10 seconds) and were asked to judge several personality traits, one of which was a scale from dominant to submissive[2]. The presented audio was recorded by several different speakers spontaneously describing a presented image. The results of this study, for the submissive-dominant trait, demonstrated that higher signal intensity increases perceptions of dominance. In this study, similar to Tusing & Dillard (2000), intensity was examined for the whole phrase and again had a more general definition of dominance, with intensity graphed in dB (100 mV for the entire statement with a 3 second averaging time).

Although these studies often do not differentiate between dominance types and generally focus on whole phrase manipulations of amplitude, the results from these studies demonstrate there is a relationship between amplitude variation and dominance ratings.

1.1. DOMINANCE TYPES. To address the differences of dominance we are interested in examining, we first need to understand what each type of dominance represents. For the purposes of our study, we used the modeled definitions of dominance proposed by Mueller and Mazur's (1997, p. 570) who provide the following description which can encapsulate this definition of *social dominance*: "a dominant person tells other people what to do, is respected, influential, and often a leader, while submissive or subordinate people are not influential or assertive and are usually directed by others." This definition was later used in a study by Puts et al. (2007) which provided the framework for the current study. Essentially, for *physical dominance*, speakers are judged by listeners whether they are more likely to be able to win a physical fight. In other words, physical dominance is a measure of physical behaviors or traits that are used to increase an individual's status within a given hierarchy, usually by influencing the behavior of others. On the other hand, *social dominance* relates to whether a speaker is more likely to be a respected leader. The associations of social dominance tie with traits often linked to connotations of *charisma*[3].

1.2. AIMS OF THE STUDY AND HYPOTHESES. As described, previous research investigated amplitude perception and dominance ratings. However, there are no studies, to our

[2] The other traits were (a) Self-doubting versus Self-confident, (b) Extraverted versus Introverted, (c) Kind versus Cruel, (d) Bold versus Cautious, (e) Lazy vesus Energetic, (j) Sociable versus Unsociable, (g) Humorous versus Serious, (h) Mature versus Immature, (i) Submissive versus Dominant, (j) Emotional versus Unemotional.

[3] The term *charisma* itself is variable with varying connotations and interpretations across individuals and groups. Previous studies (Rosenberg and Hirschberg, 2009; Tskhay, 2018) have formed the concept of *charisma* based on rating measurements of various traits. The traits often associated with charisma were enthusiasm, charm, persuasiveness, and convincingness.

knowledge, that simultaneously examine the perceptual effects of various amplitude changes for different environments like sentence level versus syllable level and how it interacts with different types of dominance (social or physical), let alone the combination of the two. Thus, the current study investigates how the produced amplitude variation on different linguistic/phonetic levels affects listeners social and physical dominance perception of these same speakers, and which of these dominance types would be most salient in high (positive) versus low (negative) ratings in listeners' perception. Three different linguistic/phonetic levels were chosen to cover a wide range of possible amplitude variation seen in acoustic phonetics and linguistics in general: sentence-level (i.e. amplitude change over the complete sentence), word-level (i.e. amplitude change for a focus word) and syllable-level (i.e. amplitude changes for either stressed or unstressed syllables).

We have the following hypotheses: Firstly, we expected to see positive correlations between increased amplitude for the whole sentence for both *social* and *physical* dominance, as was demonstrated in the previous research (Tusing & Dillard, 2000; Aronovitch 1976). We also expect to see a positive correlation between increased amplitude in word stress for *social* dominance as was found in the research by Kimble & Seidel (1991). Based on the results from these previously mentioned studies, we might also see some influence of amplitude variation in syllable stress conditions, specifically an amplitude-enhanced condition, but we expect no effect for an amplitude-reducing condition. To our knowledge, no studies on dominance and amplitude have examined gender of either the participating listener or the speaker to be judged, but we hypothesize gender of both the listener and speaker will play a role in either social or physical dominance judgments.

2. Materials and Methods

2.1 Stimuli. As speech material to be judged we selected two paragraphs and 4 isolated sentences. The paragraphs consisted of multiple simple sentences as the basis of the acoustic recordings to be used in the perception study. The paragraphs were constructed to have a neutral valence, meaning they have neither positive or negative connotations, to prevent any effect from either positive or negative valence connotation in listener interpretations of the voice. Each paragraph was approximately 12 seconds long. The paragraphs were:

i) There is a house on the street and the kitchen door is open. Inside the kitchen, there's some tablecloths in a basket. A spoon is on the table beside a coffee cup. I see a rug on the floor and magnets on the fridge.

ii) The bedroom has two windows and a closet. A painting is hanging on the wall beside a clock. A dresser is across from the bed. Four drawers are in the dresser. There is a book and a lamp on the nightstand.

The following sentences were used:

i) There are some table clothes in a basket.
ii) They drink tea or coffee in the afternoon.
iii) Four drawers are in the dresser.

iv) It is over forty miles to the nearest town.

Six native Canadian English speakers (3 female, 3 male) recorded the paragraphs. Of these speakers, four were professional voice actors (2 male, 2 female), and two were Linguistics graduate students at McMaster University (1 female, 1 male). Due to lockdown restrictions associated with COVID-19 the four professional voice actors used their own high-quality microphones and adequate recording environments to record their productions and were directed and monitored via Zoom by the authors of this study. The two graduate student speakers recorded the stimuli using a high-quality microphone (Rode NT1A) and Focusrite Scarlet audio interface inside the soundproof booth of the Phonetics Lab at McMaster University. For all recordings, microphone distance was specified to be around 10 cm with the microphone being horizontally off-centre (from the lips) by around 30 to 45 degrees.

Both paragraphs and all four sentences were repeated 3 times. The best of these 3 productions was then selected as perceptual stimulus for the listeners. Prosodic differences were as tightly controlled as possible across speakers and conditions.

After recording, the audio samples were screened with the audio editor Amadeus Pro (Hairer, 2021) and carefully checked for achieved accuracy and consistency of produced vocal intensities, the lack of undesired voice quality changes (e.g. sentence-final glottalization) and differences in articulation (e.g. different speech rates and hyper/hypoarticulation) by the two authors of this study. Additionally, a steep high pass filter (80 Hz for male speakers and 150 Hz for female speakers; 24dB/octave linear-phase) was applied to remove and attenuate any additional low frequency noise which may have been a part of the original recordings. Following the filtering, as final step before amplitude manipulations we conducted a loudness normalization of all stimuli. For this normalization all stimuli (i.e. all paragraphs and sentences for all speakers) were loudness-normalized to 65 dB using Praat, a computer program for phonetic analysis and manipulation, (Boersma & Weenink, 2022) intensity normalization procedure.

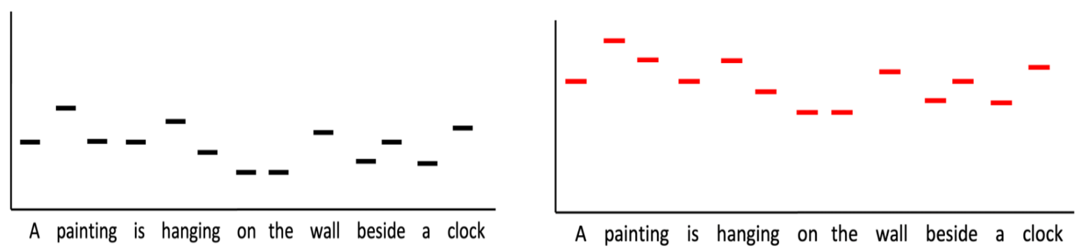
2.2 AMPLITUDE MANIPULATIONS. Following the described recording procedure, each sentence/paragraph was then manipulated in amplitude on three levels: (1) sentence stress, (2) word stress, also referred to as focus in linguistics and (3) syllable stress (stressed versus unstressed syllable amplitudes). Each of the three manipulation levels was performed on the baseline audio file (i.e. the speaker-produced audio file for each speaker and each paragraph/sentence) in Adobe Audition (Adobe Inc., 2021), resulting in an *enhanced condition* (i.e. the *enhanced* condition had increased amplitude values on either sentence, word or syllable level). Additionally, for the syllable stress condition we also included a *reduced* condition. The actual amplitude manipulations will be explained in the following paragraphs.

2.2.1. PHRASAL OR SENTENCE STRESS. Compared to the baseline condition, sentence stress was manipulated for the entire sentence or paragraph reflecting overall uniform intensity differences from the beginning until the end of the sentence or paragraph. The overall sentence intensity was increased by 6 dB, meaning the whole sentence had an increase in volume and thus perceived loudness. In other words, we increased the intensity on the sentence- or paragraph-level and compared these manipulations to the originally produced utterance (i.e. baseline) to examine how the perception of that speaker in a higher intensity condition compares to the perception of the same speaker producing exactly the same sentence/paragraph, but with normal (i.e. -6dB) intensity as baseline.

To help conceptualize this concept we present here a visualization: In Figure 1, the exemplary sentence "A painting is hanging on the wall beside a clock", is presented on the x-axis and the amplitude levels for each syllable of the sentence are stylized on the y-axis. The horizontal lines in the chart represented the amplitude levels for each syllable throughout the utterance. The right graph in Figure one uses red horizontal lines to show the performed manipulation, i.e the shift upwards in intensity to reflect overall intensity increases for all syllables in that sentence (+6dB). Please note that for the perception experiment a direct comparison of a possible listener effect is only possible by directly comparing the ratings for the enhanced version (the right graph in Figure 1) with respect to the presented baseline version (the left graph in Figure 1).

Figure 1

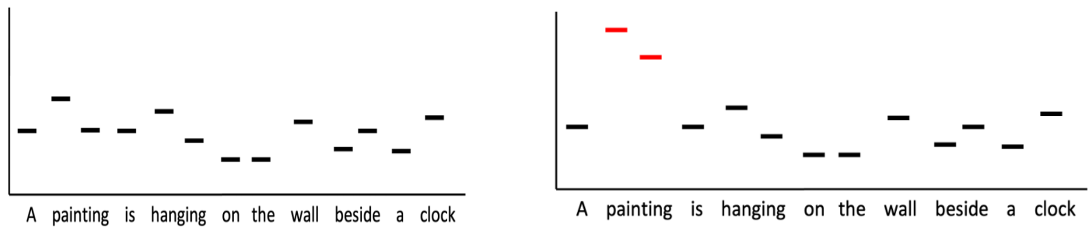
Example sentence for the baseline intensity condition (left) and sentence intensity increase (right)



2.2.2. WORD STRESS OR FOCUS. Word stress, also known linguistically as focus, was manipulated by increasing the amplitude of one specific word of a given sentence. One singular focus word of *each sentence*, including all sentences of a given paragraph, was increased in intensity by 6dB. This means that now one word is in a stressed or focus position in that sentence, with respect to intensity differences[4]. The focus word which was increased for each sentence was determined before manipulation and was kept consistent across all speakers. Figure 2 (on the right) demonstrates this manipulation. The red horizontal lines, again, demonstrate the stimulus manipulation, the baseline of this sentence would be the representation in Figure 1. The red amplitude lines of the focus word "painting" have shifted up in intensity, representing an increase in overall intensity of this word (6dB).

[4] But please note that other acoustic parameters to signal word focus or stress are not manipulated here: fundamental frequency or pitch differences, duration differences, and formant or articulation differences of stressed versus unstressed syllables.

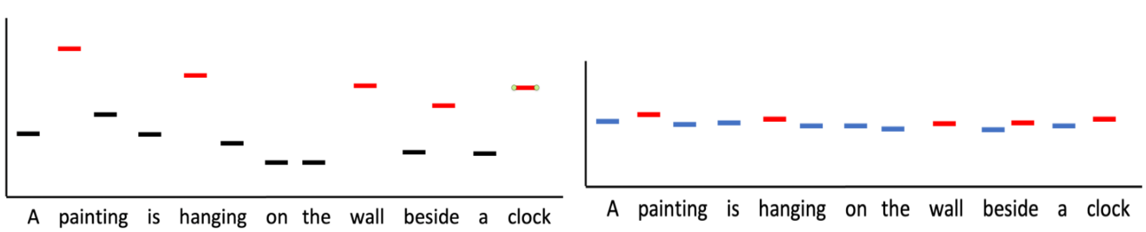
Figure 2
Example of word/focus intensity increase (right panel) compared to the baseline intensity condition (left panel).



2.2.3. Syllable stress. Lastly, syllable stress intensity, or more precisely the intensity difference between stressed and unstressed syllables in a given sentence, was manipulated in two ways: *enhanced* (as in the previous two conditions) and additionally *reduced*. These manipulations were conducted to investigate the effect of varying intensity differences contrasting stressed/unstressed syllables on listener dominance ratings.

2.2.3.1. Enhanced Syllable stress. The first variation aimed to enhance syllable stress. Here the intensity of all stressed syllables in a sentence/paragraph was increased by 6 dB, compared to all unstressed syllables. Figure 3 demonstrates how this manipulation was implemented. For each stressed syllable in the sentence, it can be seen the red horizontal lines have been shifted upward in intensity on the left sentence (again, compared to the baseline in Figure 1), reflecting this manipulation.

Figure 3
Example of syllable intensity increase (left). Example of syllable intensity reduction (right). Please note that increases of stressed syllable amplitudes are marked in red, decreases of unstressed syllable amplitude are marked in blue.



2.2.3.2. Reduced Syllable stress. The second variation was a reduction environment in which all audio stimuli had both the intensity of stressed syllables reduced *and* the intensity of unstressed syllables increased. This was done to create an effect of a speaker with reduced distinction of stressed versus unstressed syllable intensity in his/her speech production. We carefully contemplated whether (1) only the decrease of the stressed syllable intensity (to get close to the intensity of the unstressed syllables) or (2) only the increase of the unstressed syllable intensity (to get close to the intensity of the stressed syllables) or (3) both an decrease of stressed syllables intensity and increase of unstressed syllable intensity would be the best way to achieve the aim of an indistinct speaker (with respect to syllable intensity). A number of different manipulation examples showed that option (3) achieved the highest naturalness and sounded the most convincing, so we continued with the manipulation using the third option. Due to the variation across speakers of naturally produced stress these manipulations had to be done individually and manually. Figure 3 (on the right) illustrates how a combination of decreasing the intensity of stressed syllables plus increasing the intensity of unstressed syllables resulted in a flattened intensity contour over the complete sentence.

After all manipulations for each condition were applied, amplitude normalization to 65 dB was then preformed again for word stress and syllable stress conditions, minus the baseline conditions which were previously normalized to 65 dB. Sentence stress was not normalized as this would have resulted in identical amplitude measurements to those of the baseline condition (i.e., for sentence stress, the overall sentence intensity was just increased by 6 dB). The final stimuli count was 180 acoustic stimuli (6 speakers x 6 different language materials [2 paragraphs, 4 sentences] x 5 manipulations (baseline, 1 reduced condition, 3 enhanced conditions). Stimuli were not repeated, so each acoustic stimulus was only played once for each set of questions.

2.3. PARTICIPANTS. 71 participants took part in the perception study. Participants were recruited from Prolific, an online participant recruitment platform, as well as other online platforms such as Facebook. All participants reported normal hearing and cognition. They answered a set of demographic questions including gender, age, acquired and spoken languages and musical education background (question: "Do you play an instrument?" and "If yes, please specify which instrument and for how long?"). Due to restrictions for in-person experimentation, the experiment was conducted online using participants' personal laptops with the use of their own headphones. The duration of the experiment was around 60 minutes, including the pre-screening component. Participants were financially compensated. Ethics clearance for this study was obtained from the McMaster Ethics Board (MREB).

2.4. EXPERIMENTAL SETUP

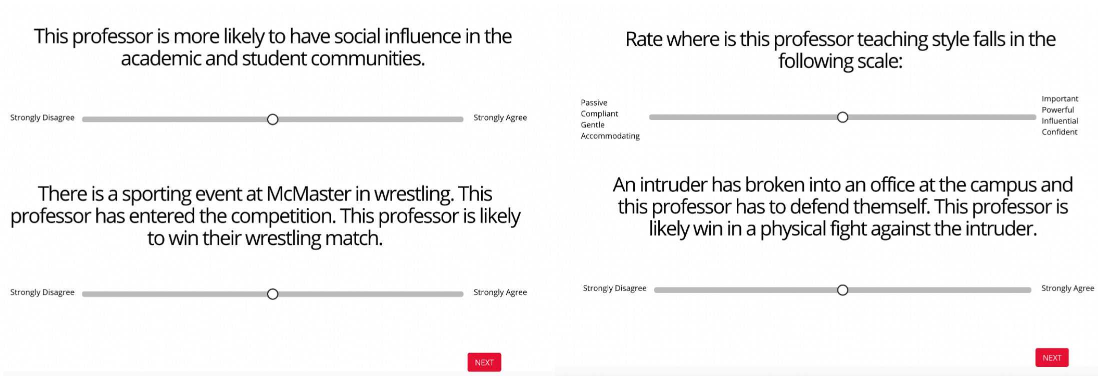
2.4.1. SCALES. One effective way of eliciting the perception of personality traits is the use of continuous sliding scales. Several studies (Puts et al., 2007; Rosenberg & Hirschberg 2009; Berger et al., 2017; Niebuhr et al., 2018) have used various forms of sliding scales. The general structure is a statement or question (eg. "The speaker is X") with a scale to opposing end scales (e.g. *strongly agree* on one end to *strongly disagree* on the other) where participants select a position on the scale according to their judgement of the stimuli in relation to the statement or question. Using a variation of the sliding scales within the study by Puts et al. (2007) as well as an additional 2 questions dominance questions (one for physical dominance and one for social dominance), the current study consisted of a total of 4 statements with scales (2 physical dominance questions and 2 social dominance questions).

2.4.2. PRESENTED STATEMENTS. The study was focused on an academic setting regarding professors and thus structured around the dominance ratings for these professors. The statements modeled from the Puts et al. (2007) study asked listeners to judge if a speaker was more likely to win a physical fight (classified as physical dominance; endpoints of “strongly disagree” to “strongly agree”), and whether the listener was “extremely dominant” or “extremely submissive” (social dominance). As previously mentioned, in addition to these 2 statements, we created and added two additional scales (one for social and one for physical). These statements can be seen below.

Participants were presented acoustic stimuli through headphones containing either one paragraph or one sentence recorded by one of the 6 individual speakers. Simultaneously to the audio stimuli, participants were presented a visual screen which contained 2 statements with a sliding scale; one statement focusing on social dominance and the other one focusing on physical dominance. On a following screen, 2 different statements were presented (again, one social and one physical dominance) were presented along with the repetition of the exact same stimuli presented on the previous screen (see Figure 6). Participants were asked to listen to the audio stimuli and rate the statements using a continuous sliding scale, based on their own perceptions of the presented speech. Thus, these ratings were based on a duplet of two statements, relating to whether the speaker was likely to win in a physical fight (physical dominance) or whether a speaker was a respected leader (social dominance). Essentially, the responses gave a rating on a scale from 0% (strongly disagree) to 100% (strongly agree).

Figure 4

Participant trial screens. Please note that the two statements on the left are an example of the first presented screen, the two statements on the right are an example of the second screen.



The *physical* dominance statements were:

i) There is a sporting event at McMaster in wrestling. This professor has entered the competition. This professor is likely to win their wrestling match. (strongly disagree to strongly agree)

ii) An intruder has broken into an office at the campus and this professor has to defend themself. This professor to win in a physical fight against the intruder. (strongly disagree to strongly agree)

The *social* dominance statements were:

i) Rate where is this professor teaching style falls in the following scale (one end of scale: passive, compliant, gentle, accommodating, other end of scale: important, powerful, influential, confident)

ii) This professor is more likely to have social influence in the academic and student communities. (strongly disagree to strongly agree)

2.5. STATISTICAL ANALYSIS. The statistical analysis was performed using the software R (R Core Team, 2018) and RStudio (R Studio Team 2018). Parametric (e.g. ANOVA) or non-parametric (e.g. Wilcoxon) tests (depending on tests for normality of the data distributions) were used to determine whether listener responses/scores for each examined intensity environment as a dependent variable would be significantly different compared to the baseline condition.

3. RESULTS & DISCUSSION. Figure 5 shows the mean and standard errors for all examined test conditions for the three linguistic levels (*sentence*, *word* and *syllable* level). Shapiro-Wilk tests for normality performed on the data was significant, so we cannot assume a normal distribution of our dataset and thus decided to conduct significance tests using the non-parametric Wilcoxon signed-rank (matched sample) test. We conducted pairwise comparisons to determine the statistical significance of each tested variable (enhanced sentence stress, enhanced word stress, enhanced syllable stress and so on), each compared to the baseline production (i.e. no enhancement/reduction). As a result, 4 pairwise Wilcoxon tests were conducted, the significance values are shown in Table 1 and Bonferroni-correction was applied to the significance levels.

Figure 5

Dominance ratings for all listeners aggregated over all 6 speakers' productions, split by physical versus social dominance. The y-axis displays the slider position percentage with 0% corresponding to the left extreme value and 100% as the right extreme value of the slider. The x-axis shows the different amplitude manipulations levels (sentence, syllable, word)

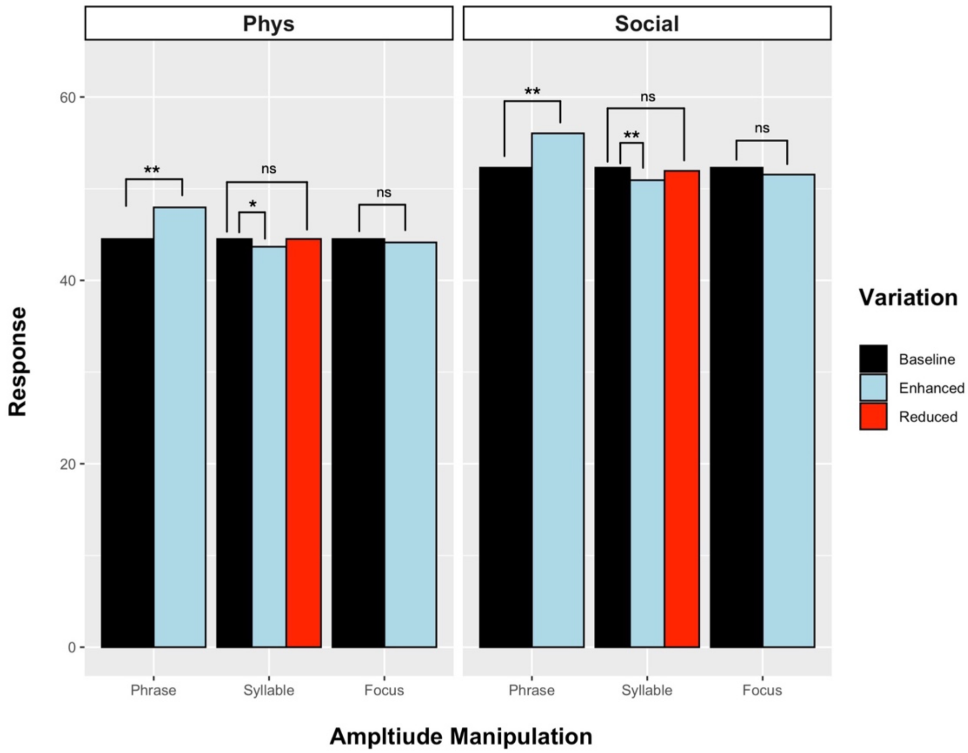


Table 1
Results of the Wilcoxon signed-rank pairwise significance test (z-values and p-values) for each examined amplitude variation by dominance type. Bonferroni correction was applied (due to multiple comparisons), the traditional p = .05 threshold is thus now at p=.0125 (and p = .01 results in p = 0.0025).

	Physical Dominance		**Social Dominance**	
	z-score	p-value	z-score	p-value
Enhanced Sentence Stress* vs *baseline	-13.822	**p < .0025 ****	-13.649	**p < .0025 ****
Enhanced Word Stress* vs *baseline	-0.691	p = 0.489	-2.357	p = 0.018
Enhanced Syllable Stress* vs *baseline	-2.634	**p = 0.008 ***	-4.405	**p < .0025 ****
Reduced Syllable Stress* vs *baseline	-0.029	p = 0.976	-1.646	p = 0.099

When examining the data, it can be seen that two linguistic levels, the *sentence* level and the *syllable* level, resulted in significantly different dominance ratings for both *physical* and *social* dominance in the enhanced condition. The intermediate *word* level manipulation did not result in significantly different dominance ratings. Furthermore, figure 5 shows that the direction of the dominance rating differences comparing manipulated versus baseline condition changes from sentence to syllable level: Whereas on sentence level participants showed *higher* dominance ratings for the *enhanced* manipulation, for syllables (and also for words) the *baseline* condition was rated *higher* (and thus more dominant) than the manipulated *enhanced* condition. Please note however that for word level this difference was not statistically significant. Finally, the reduced manipulation condition for syllables did not result in statistically significant dominance ratings compared to the baseline.

Our results of statistical significant differences for the high-level *sentence* condition confirm the results of previous research: increasing the amplitude increases the perceptual ratings of dominance in a speaker on sentence level (Tusing & Dillard, 2000; Aronovitch, 1976). As can be seen in Figure 5, this is true for both types of dominance, *social* and *physical*. Furthermore, our results contradict with the results by Kimble & Seidel (1991) who found *word* stress amplitude increases result in high scores within social dominance environments. In our data there were no significant differences on the intermediate *word* level.

With respect to the low-level *syllable* conditions we are not aware of previous literature examining intensity differences on that level, so our results shed light to the question how syllabic intensity manipulation differences affect dominance ratings of listeners: For both *social* and *physical* dominance, enhancing intensity differences between stressed and unstressed *syllables* leads to significant dominance rating differences, however differently than expected: Surprisingly the *enhanced* condition lead to significantly lower dominance ratings compared to the baseline condition, whereas the *reduction* condition did not significantly influence listeners' ratings. It should also be noted that the effect size of the *syllable* level condition is much smaller compared to the high-level *sentence* condition, so it appears that intensity changes on sentence/phrase level are driving more salient and robust changes compared to low-level syllabic changes.

3.1. INFLUENCES OF GENDER. We were also interested to examine which influence the variable gender would have on participants' dominance ratings. For our data, two types of gender influences can be examined: speaker gender, i.e. the influence of the gender of the *speaker* producing the stimuli, and participant gender, i.e. the influence of the gender of the *listener* judging the stimuli. Table 2 presents the significance values comparing male versus female speakers (left column) or participants (right column), split by *physical* or *social* dominance ratings. The values used for these Wilcoxon comparisons were the difference scores between test condition (e.g. enhanced syllable ratings) minus the baseline condition, thus providing the pure effect of intensity manipulation change with respect to the judged baseline. It can be seen in the table that only one condition shows significant gender differences: Female participants judged the enhanced *syllable* condition for *physical* dominance significantly different from male listeners, suggesting that enhancing syllable stress intensity differences plays a role in listeners' dominance ratings for physical dominance only. All other gender comparisons were not significant, for both speaker and listener comparisons.

Table 2
Results of the Wilcoxon signed-rank pairwise significance test (p-values) comparing the effects of speaker and participant gender (male versus female) for physical and social dominance ratings. Bonferroni correction is applied for significance thresholds.

	Speaker Gender		**Participant Gender**	
	Physical	Social	Physical	Social
Enhanced Sentence Stress	p = 0.311	p = 0.2511	p = 0.1413	p = 0.144
Enhanced Syllable Stress	p = 0.1867	p = 0.2039	**p = 0.0121 ***	p = 0.5727
Reduced Syllable Stress	p = 0.377	p = 0.9702	p = 0.3134	p = 0.445
Enhanced Word Stress	p = 0.3273	p = 0.2808	p = 0.9003	p = 0.81

4. CONCLUSIONS. Our results demonstrate that increasing intensity on sentence/phrase level leads to statistically significant differences for both *physical* and *social* dominance. It is the also the most salient of all examined conditions. Our results thus confirm previous research examining the relationship of whole-sentence intensity increases and dominance ratings: the louder the sentence the greater the perception of dominance for both social and physical dominance (Tusing & Dillard, 2000; Aronovitch, 1976). Although these results, on the first view, may suggest that the louder the production of a phrase/sentence the greater the perceived dominance - without a lower amplitude baseline condition to serve as comparison the increased amplitude variation cannot be really applied to real life conditions. In other words, just increasing all full sentence amplitudes as a speaker in order to be perceived as more dominant by listeners will not work in real life, since the comparing baseline of a lower amplitude condition is missing (and also most importantly the need for identical sentence material to be judged).

One other aspect is that other studies, to our knowledge, have not investigated gender as an influencing factor for dominance ratings. As our results demonstrate there is no effect of gender on any of the examined conditions for *speakers*. For *listener* gender however, enhancing *syllable* stress amplitude, but for *physical* dominance ratings only, showed statistically significant differences comparing female and male listeners. It should be noted that all other comparisons were not significant. These results regarding gender confirms, at least partially for the level of *syllable* stress for *physical* dominance, our hypothesis that gender could play a significant role in dominance ratings in the paradigm of amplitude manipulations, but only for the examination of *listener* gender, not *speaker* gender. Explanations towards the rationale can remain only speculative at this point of research. It is possible that it is linked to intergroup connotations regarding dominance, with males gravitating towards more *physical* associations of dominance while women prefer a more *social* based orientation. However why these gender differences occur on *syllable* but not also on *sentence* level is not clear. More research is needed to examine this relationship further.

REFERENCES

ADOBE INC. (2021). *Adobe Audition*. Retrieved from https://adobe.com/products/audition

ARONOVITCH, C. D. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. Journal of Social Psychology, 99, 207-220.

BERGER, S., NIEBUHR, O., & PETERS, B. (2017, March). Winning over an audience–A perception-based analysis of prosodic features of charismatic speech. In *Proc. 43rd Annual Conference of the German Acoustical Society, Kiel, Germany* (pp. 1454-1457).

BULLER, D. B., & BURGOON, J. K. (1986). The effects of vocalics and nonverbal sensitivity on compliance: A replication and extension. Human Communication Research, 13, 126-144.

BULLER, D. B., & AUNE, R. K. (1988). The effects of vocalics and nonverbal sensitivity on compliance: A speech accommodation theory explanation. *Human Communication Research*, *14*(3), 301-332.

HAIRER, M. (2021) Amadeus Pro (v. 2.8.8) [software]. London, United Kingdom; HairerSoft.

HARRIGAN, J. A., GRAMATA, J. F., LUCIC, K. S., & MARGOLIS, C. (1989). It's how you say it: Physicians' vocal behavior. *Social Science & Medicine*, *28*(1), 87-92.

KIMBLE, C. E., & SEIDEL, S. D. (1991). Vocal signs of confidence. *Journal of Nonverbal Behavior*, *15*(2), 99-105.

MUELLER, U., & MAZUR, A. (1997). Facial dominance in Homo sapiens as honest signaling of male quality. Behavioral Ecology, 8, 569–579.

NIEBUHR, O., & FISCHER, K. (2019, September). Do not Hesitate!-Unless You Do it Shortly or Nasally: How the Phonetics of Filled Pauses Determine Their Subjective Frequency and Perceived Speaker Performance. In *INTERSPEECH* (pp. 544-548).

NIEBUHR, O., SKARNITZL, R., & TYLEČKOVÁ, L. (2018, June). The acoustic fingerprint of a charismatic voice-Initial evidence from correlations between long-term spectral features and listener ratings. In *Proceedings of Speech Prosody* (Vol. 9, pp. 359-363).

BOERSMA, P, & WEENINK, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.23, retrieved 8 October 2022 from http://www.praat.org/

PUTS, D. A., HODGES, C. R., CÁRDENAS, R. A., & GAULIN, S. J. (2007). Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), 340-344.

QUENÉ, H., BOOMSMA, G., & VAN ERNING, R. (2016). Attractiveness of male speakers: effects of voice pitch and of speech tempo. In *Proceedings of Speech Prosody* (Vol. 8, pp. 1086-1089).

RAFIKOVA, A., VALUEVA, E., & PANFILOVA, A. (2022). Voice and Psychological Characteristics: A Contemporary Review. *Psychology. Journal of Higher School of Economics*, *19*, 195-215.

R CORE TEAM. (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing,* Vienna, Austria. Retrieved from https://www.R-project.org/.

ROSENBERG, A., & HIRSCHBERG, J. (2009). Charisma perception from text and speech. *Speech Communication*, *51*(7), 640-655.
RSTUDIO TEAM (2022). *RStudio: Integrated Development for R.* RStudio, PBC, Boston, MA Retrieved from http://www.rstudio.com/.
SCHERER, K. R., & LONDON, H., & WOLF, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. Journal of Research in Personality, 7, 31-4
SCHERER, K. R. (1974). Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences. In S. Weitz (Ed.), Nonverbal communication (pp. 105-111). New York: Oxford University Press.
TUSING, K. J., & DILLARD, J. P. (2000). The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence. *Human Communication Research*, *26*(1), 148-171.
YANG, Z., HUYNH, J., TABATA, R., CESTERO, N., AHARONI, T., & HIRSCHBERG, J. (2020). What makes a speaker charismatic? producing and perceiving charismatic speech. In *Proc. 10th International Conference on Speech Prosody* (Vol. 2020, pp. 685-689).

CHAPTER 3

The Effects of Different Voice Qualities on the Perceived Personality of a Speaker

This study has been submitted and published in *Frontier's in Communication: Psychology of Language* as Pearsell, S., Pape, D. (2023). The effects of different voice qualities on the perceived personality of a speaker.

Abstract

Although previous studies investigated various aspects of voice quality perception and personality attribution there are no studies, to our knowledge, which simultaneously examine and compare the perception of various voice qualities when produced by the same individual. This work investigates how laryngeal and supralaryngeal voice quality variations of a speaker affect *listeners'* perceived personality traits (and thus perceived charisma) of that same speaker. Six Canadian English speakers produced paragraphs varying the following voice qualities: *modal*, *creaky*, *breathy* (natural and artificial), (hyper-)*nasalization*, and *smiling* (natural and extreme). Listeners of a perception experiment were then tasked to rate ten statements for each presented audio stimulus. Statements were selected corresponding to a sub-section of the Big 5 personality traits shown to be linked to charisma perception. Results show significantly more positive listener ratings (i.e. higher ratings compared to *modal*) with medium effects sizes for both *smiling* variants across all personality traits. In contrast, *creaky* was perceived significantly more negatively overall for all personality traits, with a medium effect size. *Nasal* and *breathy* still achieved statistically significant rating differences compared to the *modal* baseline. However, the overall effect pattern was more complex, and effect sizes were small or negligible. Additionally, we found consistent differences for some voice qualities when examining listener ratings comparing male versus female *speakers*: for both *creaky* and *smiling* (but not for other voice qualities), female speakers were rated more negatively when producing *creaky* for some personality traits, whereas both *smiling* variants were consistently rated higher for females compared to males.

1 Introduction

Recently there has been an increased interest in analyzing and understanding the effects of speakers' voice quality differences and how these *produced* differences may impact the *perception* of these speakers' personality traits. Areas of interest in this topic range from clinical techniques for best practices for a healthy vocal production (while avoiding vocal strain) to popular culture tips to sound more professional or speak more effectively. One specific area of interest is the role of voice quality in the perception of a speaker's personality traits, and, more generally, how these traits relate to the perception of speaker *charisma.*

1.1 Voice Quality versus vocal quality

Voice quality has been defined as "the quasi-permanent quality of a speaker's voice" (Abercrombie, 1967) and "those characteristics which are present more or less all the time that a person is talking. It is a quasi-permanent quality running through all the sound that issues from his mouth" (Abercrombie, 1967). Following this definition, and in line with researchers like Laver (1980) and Esling et al. (2019), voice quality differences are based on the specific auditory coloring of an individual's voice as a result of the variations of both *laryngeal* and *supralaryngeal* features which continuously occur throughout an individual's speech production. Several significant factors play a role for the variation of different laryngeal voice qualities: sub- glottal pressure (the air pressure below the vocal folds), medial compression (the contraction of the lateral cricoarytenoid muscles causing adduction; how tightly the vocal folds are pressed together), adductive tension (how tightly the arytenoid cartilages are pressed together at the posterior end of vocal folds) and longitudinal tension [the tension or slack of vocalis, thyroid and cricoid muscles, as well as the cricothyroid muscles (Laver, 1980)]. Following Laver's research, the most common phonation types, or laryngeal settings, are (i) *modal* or *normal voice*, the baseline (and non-pathological) voice setting, (ii) *breathy voice,* which has a high rate of air flow during production, (iii) *creaky voice* (also known as *vocal fry, laryngealization* or *glottalization*) characterized by very low frequencies which can be irregularly timed, (iv) *harsh voice*, a speech pattern with a normal fundamental frequency but aperiodicity or noise in the spectrum, and lastly (v) *tense or strained voice*, produced with a low rate of air-flow (often described as a "metallic" voice).

To conceptualize voice quality, it is helpful to think of each voice quality as a landmark on a continuum, with *breathy* on one end of that continuum (produced with a more open glottis), and *creaky* on the other end (produced with a constricted glottis). *Modal* voice is found between these two extremes. As a general notion of *modal* voice, this vocal quality has a more regular and periodic vibration pattern; there is no audible friction of the vocal folds, and the muscular tension is moderate. The vibrations are regular along of the vocal folds, often characterized as a "neutral mode of phonation" (Laver, 1980, p. 110). Medial compression, adductive tension, and airflow from the lungs are all moderate, and the longitudinal tension is low (vocal folds are shorter and thicker). The described voice quality landmarks vary slightly between individuals but maintain the same directional proximity to one another (*breathy* on one end, *modal* more central, and *creaky* on the other end).

Although these laryngeal features are the most dominant aspect in the description of different voice qualities, both the Abercrombie (1967) and Laver (1980) frameworks include suprasegmental modification of non-laryngeal features such as retroflexion/retraction, smiling or nasality. In this paper, we adapt these definitions of voice quality, which are also supported by the ANSI definition (i.e., that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar). Still, throughout this paper, we will use the term *vocal quality* to stress that this term would include both laryngeal features (e.g., modal, creaky, and breathy voice qualities) but also consistent and continuous suprasegmental feature variation (such as smiling and nasality, found on supralaryngeal and suprasegmental level). We hope that by using the term *vocal quality* we clearly define the inclusion of non-laryngeal vocal tract features since the term *voice quality* is often used very differently in the literature[5]. Please note that the suprasegmental features of interest in the current study, smiling and hypernasality, can of course co-occur with laryngeal voice quality variations. For the purposes of this study, the suprasegmental features of interest are produced with underlying modal voice quality, with modal voice representing the baseline measurement for each speaker.

1.2 Vocal quality and personality perception

Previous research has examined various acoustic features and perceptual cues and their relationship to personality trait attribution. Some of these studies have investigated the relationship between independent features of segmented speech signals such as f0 and pitch[6] (Puts et al., 2007; Rosenberg and Hirschberg, 2009; Quené et al., 2016; Berger et al., 2017), nasality and filled pauses (Möbius, 2003; Niebuhr and Fischer, 2019), amplitude or loudness (Novák-Tót et al., 2017), harmonics frequencies (Collins, 2000; Hodges-Simeon et al., 2010), and vocal quality (Wolf, 2015; Abdelli-Beruh et al., 2014) alongside their interaction with various personality traits. These results suggest that individual variation within physiological aspects of speech can play an important perceptual role in personality trait ascription.

In earlier research on the perception of vocal quality and perceived personality attributes, Pittam (1985) examined different vocal qualities of speakers and the impact of these qualities on listeners' ratings of solidarity, attractiveness, and status of the speaker. This study found that listeners' ratings of solidarity with a speaker were greater when there was the presence of either *breathiness* or *whispery*[7] qualities in

[5] Our decision to use vocal quality instead of voice quality for this paper stems from discussions with other researchers who often defined voice quality as purely consisting of laryngeal differences.
[6] Pitch is a perceptual term taking into account the different acoustic properties of a complex acoustic waveform (normally consisting of the fundamental frequency and a number of optional harmonics). Therefore, pitch perception values can be different from measured fundamental frequency values.
[7] Whispery voice is categorized as a combination of glottal friction and voicing. This combination creates greater amounts of inter-harmonic noise, creating an almost flat spectrum with increased levels

the speaker's voice. Perceptions of status were higher for *tense* voices as well as *breathy* voices compared to *whispery* and *nasal* voices (Pittam, 1985). In another study, Laver (1972) demonstrated an association of *breathy* voice with perceived higher sexuality and sensuality when the speaker was female but not when the speaker producing that *breathy* quality was male. Other studies have also demonstrated a correlation between certain vocal qualities and perceptions: the more significant the creakiness of a speaker, the higher the perceptions of that speaker's dominance or higher social status; the harsher the voice quality, the lower the perception of prestigious status (Esling, 1978; Scherer, 1979). Additionally, participants (who were described as young adults) rated voices with increased *creakiness*, above all the other vocal qualities assessed, as older. Esling (1978) and Scherer (1979) also suggest that this perception of age, as a result of the presence of *creakiness* in vocal quality, may account for the decrease in ratings associated with the friendliness and attractiveness of a speaker.

One major theory of personality and its associated traits is the *Big 5 of Personality Traits* (Norman, 1963; McCrae and John, 1992). Within this theory, personality traits are categorized and defined within five groups: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (or OCEAN for short). The framework for our personality trait perception is based on this theory resulting in a broader categorization of attributes such as attractiveness or speaker status into one or more of these five personality traits. Despite the interest in vocal quality and personality perception, to our knowledge, the current research remains limited to the focus of between-subject designs. That is, different voice quality conditions were always confounded with different speakers, for example, examining *creaky* vocal quality and thus only the influences of that *creaky* voice quality on listener perception. Understanding both the perception and production of multiple, potentially influential vocal qualities an individual is consistently able to produce can provide insight into many areas of interest. These interests could range from (1) understanding how listeners perceive a multitude of possible variations within a speaker's productions, as well as (2) clinical opportunities for those suffering from pathologies impacting their productions, (3) to professional opportunities for those who are outside academia to improve the effectiveness of their speech productions and understanding how their voice and its productions are perceived by an audience. Furthermore, currently it remains unclear what aspects of vocal quality variation are most salient for the concept of a *charismatic* speaker (Signorello and Demolin, 2013).

The present study investigates vocal qualities varied in a within-subject design, focusing on the following vocal qualities: *modal*, *breathy*, *creaky*, representing opposing ends of the voice quality spectrum as well as a medial point between the two, and the additional qualities of *nasality* (specifically hypernasality), and *smiling*. Within this study, these vocal qualities are rated explicitly in terms of within-subject personality traits and more specifically in terms of charisma-related traits.

---

of energy (Laver, 1980). Breathy voice differs from whispery because of weaker medial compression and a decreased degree of voicing effort. However, Laver (1980) notes the perceptual boundary is not clear between whispery and breathy. In this paper we use the term breathy quality.

1.2.1 Creaky voice

Creaky voice, also referred to as *vocal fry*, *glottal fry*, *laryngealization*, and *glottalization*, has been extensively researched. *Creaky* voice can be categorized by its irregular vocal fold vibrations created by the amalgamation of high adductive tension, low longitudinal tension, high medial compression and low subglottal pressure (Laver, 1980; Ladefoged and Johnson, 2011). It usually occurs at the lower end of a speaker's f0 range. Gick et al. (2013) explain that "In creaky voice, the vocal folds are very shortened and slackened to maximize their mass per unit length, and the IA (Inter Arytenoid) muscles are contracted to draw the arytenoid cartilages together. This action allows the vocal folds to stay together for a much longer part of the phonation cycle than in modal voicing..., only allowing a tiny burst of air to escape between long closure periods".

1.2.1.1. Creaky voice and personality perception

Previous research remains equivocal as to the perceptual influence of *creaky* voice on a speaker's personality characteristics. One study by Yuasa (2010) found favorable listener impressions for increased usage of *vocal fry*, with associations to personality traits such as professionalism, genuineness, and nonaggressiveness, as well as other positive assumptions about a speaker (e.g., higher level of education). *Creaky* voice has also been associated with worthiness, intelligence and friendliness (Pittam, 1987). However, other studies contradict these results: Anderson et al. (2014) showed that the presence of *creaky* voice, specifically in women, has the potential to negatively impact ratings of education level, competence, and trustworthiness. Gobl and Ní Chasaide (2003) found that *creaky* voice represents impressions linked to boredom and sadness. *Creaky* voice is found to be dominant in both younger male and female populations (Wolk et al., 2012; Abdelli-Beruh et al., 2014). Despite being present in both genders, research has shown when it comes perception of creaky voice, female speakers are more frequently perceived negatively compared to male speakers (Anderson et al., 2014; Wiener and Chartrand, 2014; Pointer et al., 2022). Although these studies present conflicting results, personality traits selected across studies do not equate to the same meaning or interpretation. It also be noted, regardless of personality trait mismatching across studies, that gender (and perhaps context) appears to influence the perceptual impact of *creaky* voice on listeners, therefore providing insights for the hypothesized outcomes of the current study when varying speaker and/or listener gender.

1.2.2 Breathy voice

In voices which are considered healthy (i.e., non- pathological), *breathiness* is categorized by partial adduction along the length of the vocal folds, with both the medial compression and adductive tension at low values, thus resulting in the increased escape of air (Laver, 1980; Reetz and Jongman, 2020). The amount of air escaping during phonation can cause differences in the perceived *breathiness* of a speaker's voice, with less adduction and a more gradual closing of vocal folds making the voice sound breathier (Hanson, 1997).

1.2.2.1. Breathy voice and personality perception

As previously described, *breathiness* has been shown to increase listeners' solidarity ratings[8] and perceived status (Pittam, 1985) as well as to influence perceived sexuality and sensuality for female speakers (Laver, 1980). However, research on the influence of this specific voice quality remains limited. Understanding the gap in the literature with respect to *breathy* vocal quality can provide further insight into how vocal qualities impact listeners' categorization of speakers' personality traits.

1.2.3 Nasal voice

Nasality is a vocal quality which results from nasal sound energy in the production of a speech signal. It is the result of the velopharyngeal port being either open or closed at inappropriate times or more than acceptable in a given language or dialect.

Nasal vocal quality is the acoustic result of the sustained and excessive coupling of the nasal and oral cavities during speech and can be categorized in one of two ways: hypernasality (i.e., going toward an excess of nasality) and hyponasality (i.e., going toward the absence of nasality). Hypernasality is caused by an excess of air leaking out through the nasal cavity when speaking. This results in extra (nasal) resonances in the acoustic speech stream. This type of nasality can be a result of several factors, from physiological issues, including structural problems (e.g., shortened soft palate or movement problem causing incomplete closure of the nasal cavity) to errors in sound acquisition (e.g., not learning, normally as a child, how to control the movement of air through the vocal tract cavities). Additionally, hypernasality still can have varying degrees of presence (more nasal and less nasal) and is primarily a result of both the size and status of the velopharyngeal port opening (Watterson and Emanuel, 1981; Warren et al., 1988); however, this is a separate factor from the presence or absence of nasality in speech production. Hyponasality is the opposite of hypernasality, in which not enough air can pass through the nasal cavity, resulting in a lack of nasal resonances in the speech signal as a result of a blockage or obstruction in the nasal cavity. This vocal quality is typical of the common cold (Tull, 1999).

1.2.3.1. Nasal Voice and personality perception

To our knowledge, there is no previous research on the perceptual impact of *nasality* variation (specifically hypernasality) in non-pathological voices, presenting a knowledge gap in the literature on this vocal quality and its effect on speaker perception.

---

[8] i.e., listeners' solidarity ratings with the perceived speaker, in other words the speaker currently being rated by the listener.

It is important to note that, in principle, hypernasality could be combined with other vocal qualities, such as *breathy* or *creaky* voice. In our study, we restrict our examinations to the effect of *nasality* coupled with underlying *modal* voice, thus excluding combinations with other phonation types. Furthermore, nasal coupling is *continuously* produced by means of a lowered velum throughout the full duration of a sentence/paragraph production.

1.2.4 Smiling

The physiological movements involved in *smiling* include the widening of the mouth, retraction of the lips, the lowering of the tongue dorsum, and the tendency of a speaker to lower their jaw (Shor, 1978; Erickson et al., 2009). As a result of these movements, the vocal tract shortens, therefore altering the auditory perception of a speaker through an increase in formant frequencies as well as amplitude (Tartter, 1980). Tartter found that *smiling* has an audible effect on speech, generally associated with increased positive interpretations in a *smiling* condition.

1.2.4.1. Smiling and personality perception

A study by Vazire et al. (2009) explored the impact of the speaker's sex on the interpretation of listeners' *smiling* perceptions. The outcome of the study revealed two separate affective states, one for men and one for women. For women, *smiling* was viewed as a signal of trustworthiness and indicated warmth or enthusiasm to the listener. *Smiling* in men was interpreted as a lack of self-doubt, and increased confidence and calmness. Other research has found producing speech while smiling positively impacts speech perceptions, but has ceiling effects: excessive smiling does not increase the perception of charisma when compared to moderate smiling (Tschinse et al., 2022). For the present study, the inclusion of the *smiling* condition aims to reveal the connection between the effects of a *smiling* speaker on the perception of personality traits and effectiveness as a speaker when embedded in our experimental setup. Of particular interest for the current study, similar to the findings for *creaky* voice, is the mismatch in personality trait attribution when comparing (speaker) gender. Please note that *smiling*, like hypernasality, could be combined with other voice qualities such as *breathy* or *creaky*. In our study, we will examine *smiling* only with an underlying *modal* voice.

These vocal qualities (modal, creaky, breathy, nasal, and smiling) have been examined individually and been ascribed personality trait correlates. As previously mentioned, there remains a lack of knowledge comparing these different vocal qualities, in combination, and across individual speakers. We hope to clarify the saliency of each of these vocal qualities when compared to each other, while simultaneously clarifying their interaction with respect to personality trait association.

1.4 Personality traits and charisma

The definition of charisma presented by Niebuhr and Fischer (2019) states: "*charisma is symbolic, emotional laden, and value- based communication style signaling leadership qualities such as commitment, confidence, and competence that affect followers' beliefs and behaviors in terms of motivation, inspiration, and trust."* To further understand how to conceptualize charisma and charismatic speech research has looked at listeners' perceptual ratings for speakers' voices. These ratings were obtained through a series of presented statements correlating to charisma which listeners would rate from positive to negative, depending on the statement of each scale (Rosenberg and Hirschberg, 2009; Tskhay et al., 2018). For example, Rosenberg and Hirschberg (2009) found that *charismatic speakers* were associated with the (personality) traits of being enthusiastic, charming, persuasive, and convincing, all traits which can be found and categorized within the *Big 5* (John and Srivastava, 1999). As there is increasing interest in the sources of perceived charisma and more generally influential speakers, relying on vague interpretations of *charisma* is insufficient while using only the Big 5 of personality traits is too broad. By analyzing charisma within the traits of the Big 5 a clearer and more concrete interpretation of charisma can be established. The motivation behind our research is two-fold. Firstly, quantifying charisma based on the Big 5 allows for a targeted understanding of which attributes form different trait categories in order to create the concept of charisma, while concurrently allowing for a better "big picture" interpretation of personality traits perception using the Big 5.

Although *charisma* may not be a trait in and of itself, there are still many personality traits that coincide with charismatic features of speech, as noted in a paper by Michalsky and Niebuhr (2019). As the authors point out, studies by several other researchers have demonstrated the relationship of the *Big 5* traits to charismatic speech features. Antonakis et al. (2016) implemented a training program targeted to teach charisma to managers and business leaders using a system called *Charismatic Leadership Tactics (CLTs).* The purpose of these CLTs was to make the concept of charisma more tangible to learners. Within their research, the authors demonstrate that confidence and self-assuredness are two facets which comprise charismatic speech. When examining these facets within the personality trait dimensions of the *Big 5,* these two facets fall into the *extroversion* personality trait (Costa and McCrae, 1992; John and Srivastava, 1999). Michalsky and Niebuhr (2019) also point out that the personality trait *agreeableness* relates to charismatic features, such as kindness, warmth, and development of trust while *conscientiousness* links to job performance and self-discipline (Costa and McCrae, 1992; John and Srivastava, 1999). Using just these examples, whether charisma is a personality trait in and of itself is debatable. Despite this, the traits associated with charismatic features of speech do have a relationship with personality traits and the *Big 5,* and exploration of charisma within the Big 5 traits could provide a more general concept and understanding of the interaction of charisma perception and the use of vocal quality production.

In order to determine how different vocal qualities are attributed to the perceived personality traits of a speaker as well as how personality traits relate to charisma, the concept of personality traits needs to be further defined. As briefly mentioned above, one prominent theory of personality dimensions is that of the *Big 5 of Personality Traits* (Norman, 1963; McCrae and John, 1992). In this theory, personality traits can be described and categorized into the following sets: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (or OCEAN for short). It is important to note that each of these categories is a range of extremes. For example, extroversion is on one side of the spectrum while introversion is on the other (John and Srivastava, 1999). There is a scoring system which takes participant responses to a number of questions and rates these responses as a score from high (e.g., extroversion) to low (e.g., introversion). Figure 1 provides a visual representation and brief summary of each of the five main traits as well as the traits associated with high and low scores.

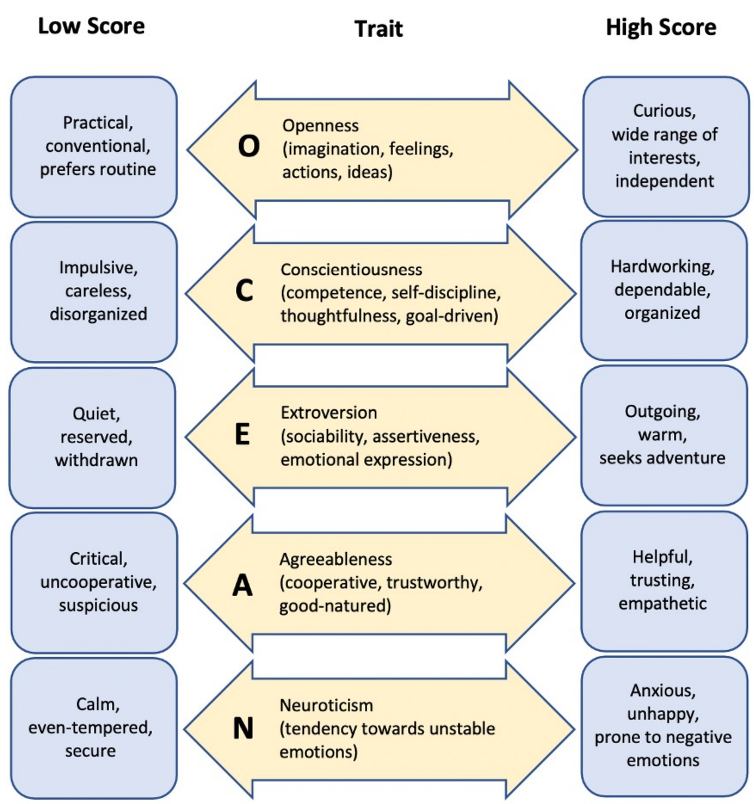


Figure 1: The Big Five Personality Traits, following Gray (2017)

The first trait is openness. This is a personality trait tied to imagination and insight, as well as openness to new experiences. Individuals with higher ratings in this trait are often perceived as more creative and have a wide-ranging set of interests. They are open, artistic, curious, and imaginative. Individuals who rate low in this trait are resistant to new ideas, are unimaginative, dislike change, and do not like to try new things (John and Srivastava, 1999). Using the questionnaires within the studies by Rosenberg and Hirschberg (2009), and Tskhay et al. (2018), we manually classified each of the questions presented there into the *Big 5 framework.* This later became the structure for our experimental design. From these two studies, the particular trait of openness has not been strongly associated as an indicator of charisma. In our experimental design, we therefore opted to omit this particular trait.

The second trait, conscientiousness, is linked to a person's attention to detail, attentiveness, and goal-directed behavior. Those with a higher score in this trait are generally categorized as efficient, organized, reliable, and responsible, while lower scores are associated with those who are less organized and more flexible in their approach to work. They may also procrastinate, lack discipline, and be careless, resulting in difficulty in completing tasks or goals (John and Srivastava, 1999). From the questionnaire list by Rosenberg and Hirschberg (2009), higher scores in conscientiousness related statements correlated to charisma but were however less proportionate in the number of statements presented than traits like extroversion and agreeableness.

Extroversion, the third trait, is related to the level and degree to which a person seeks interaction with their environment focusing on the social component. Those rating high in extroversion tend to be more social, assertive, outgoing, talkative, etc., while introversion, or those on the low rank of this trait, tend to be more reflective, and reserved, preferring solitude, avoiding being the center of attention and tend to be fatigued by an excess of social interaction (John and Srivastava, 1999). Generally, a higher rating for extroversion is characteristic for charisma perception (Vergauwe et al., 2017), and extroversion is the Big 5 personality trait that receives the highest focus when determining charismatic attribution (Rosenberg and Hirschberg, 2009).

The fourth trait, agreeableness, determines how people treat their relationships with others. Unlike extroversion, agreeableness has to do with the pursuit of relationships with motivations concentrating on people's alignment and their interactions with others (John and Srivastava, 1999). Higher ratings in this trait indicate a person who is kind, forgiving, sympathetic, and trusting. Lower rating signal skepticism, stubbornness, a lack of sympathy, and a person who doesn't care about the feelings of others. The ability to connect with people as well as develop trust are just a few aspects which have also been demonstrative of charismatic speakers (John and Srivastava, 1999). Higher agreeableness scores appear to also signal increased charisma in speakers (Rosenberg and Hirschberg, 2009).

The fifth and last trait is neuroticism. This is the trait which encompasses how an individual perceives the world, including the likelihood of inferring events as difficult or threatening as well as the inclination to experience negative emotions. People who rate high in this trait are anxious, tense, unstable, hostile, or irritable and experience dramatic shifts in mood. Those who rate lower are more emotionally stable, calmer, rarely feel sad or depressed, and do not often worry (John and Srivastava, 1999). In general, higher ratings of neuroticism have been shown to be negatively correlated with charisma and charismatic traits (Bono and Judge, 2004).

Although the Big 5 of Personality has traditionally been designed to be used by individuals based on introspection, the current study models questions/statements used in the previous studies of Rosenberg and Hirschberg (2009) and Tskhay et al.

(2018). Within these studies, questions/statements were structured to be extrospective rather than the traditional introspective structure of Big 5 questionnaires. Other research (Hart and Hare, 1994; Ziegler et al., 2010) has demonstrated that ratings given by others fall closely within the range of ratings given from introspection. Theoretically, this means that results collected from our study's extrospective structure should produce data similar to those which would have been made by introspection.

1.4 Aims of the study and hypotheses

As previously described, although there are studies which investigate individual vocal quality perception and personality attribution, there are no studies, to our knowledge, that simultaneously examine the perceptual effects of various vocal quality changes produced by the same individual speaker on the perception of charismatic traits within the context of the *Big 5*. The current research investigates how vocal quality variation (breathy, creaky, nasal, and smiling) of different speakers affects listeners' perceived personality traits and thus charisma of these same speakers. We are also interested whether one of these voice qualities is most salient in high (positive) vs. low (negative) personality trait ratings by listeners. Furthermore, we want to examine the influence of gender on listener perceptions: here we are interested in both the influence of *speaker* gender on listener ratings, but also the influence of *listener* gender.

Apart from the differences between vocal quality categories, we are also interested to examine the effects of two within- category modifications for *smiling* and for *breathy* voice. With respect to smiling, following the research by Tschinse et al. (2022) we are interested to either replicate or dispute the observed ceiling effect for *normal* vs. *extreme* smiling condition with our within-subject design, all with respect to charisma ratings. With respect to breathy voice, we aim to introduce a technical, or more artificial, noise source modification in addition to the natural speaker-produced condition, thus examining the perceptual rating difference between a naturally produced breathy voice on the one hand vs. an artificially generated (technical) breathy voice on the other hand. The motivation here is to find out whether artificial noise added to the complete communication chain (and thus *not* modulated by laryngeal differences) would influence personality trait perception. In technical terms, the technical noise should be speech-shaped to make the conditions comparable and avoid adding another confound dimension.

We have the following hypotheses:

- H1: Lower listener perception scores, or negative ratings, for *creaky* voice across all speakers (resulting in a lower rating for all investigated personality traits, including neuroticism[9]), signaling a lack of perceived charisma in speakers. Lower scores for these traits in previous studies (Bono and Judge, 2004; Rosenberg and Hirschberg, 2009; Tskhay et al., 2018) have demonstrated a correlation to negative perceptions regarding speaker charisma.

---

[9] Please see section Experimental setup for the explanation about neuroticism scores.

- H2: Higher, or more positive, listener ratings in personality traits for *smiling,* with smiling having a positive correlation with speaker charisma.
- H3 (null hypothesis): Following Tschinse et al. (2022) we expect to see a ceiling effect for *smiling*, with the natural *smiling* productions expected to have almost identical rating scores compared to the extreme *smiling* condition.
- H4 (interactional hypothesis): We predict *speaker* gender to play a role in listener ratings. Specifically, female speakers will be rated more negatively when producing *creaky* voice (i.e., receive lower personality trait scores). For male speakers, we predict a less negative (or higher score) attribution of *creaky* voice compared to female speakers, thus bringing their ratings closer to *modal* voice ratings, meaning creaky voice for female speakers would be perceived less charismatically than their male counterparts.
- H5: For naturally produced vs. technical breathy voice we expect to see perceptual rating differences, with naturally produced breathy voice ratings lower for all examined personality traits and therefore rating lower in charisma. The reason for the lower expected ratings for natural breathy voice is that we assume that listeners are able to distinguish between noise as part of the speaker's laryngeal system (and thus being constantly modified by the speaker's production), whereas a channel-induced noise source could be better separated from the judged speaker characteristics, and thus would influence personality perception ratings less than the natural breathy condition.

2 Materials and Methods

2.1 Stimuli

We selected two paragraphs consisting of multiple simple sentences as the basis of the acoustic recordings to be used in the perception study. The paragraphs were constructed to have a neutral valence to prevent any impact from positive or negative valence in listener interpretations of the voice. Each paragraph was ∼12 s long.

1) There is a house on the street and the kitchen door is open. Inside the kitchen, there's some table clothes in a basket. A spoon is on the table beside a coffee cup. I see a rug on the floor and magnets on the fridge.
2) The bedroom has two windows and a closet. A painting is hanging on the wall beside a clock. A dresser is across from the bed. Four drawers are in the dresser. There is a book and a lamp on the nightstand.

Six native Canadian English speakers (3 female, 3 male) recorded the paragraphs. Of these speakers, four were professional voice actors (2 male, 2 female), and two were Linguistics graduate students of McMaster University (1 female, 1 male). Due to lockdown restrictions associated with COVID- 19, the four professional voice actors used their own high-quality microphones and adequate recording environments to record their productions and were directed and monitored *via* Zoom by the authors of this study. The two graduate student speakers recorded the stimuli using a high-quality microphone (Rode NT1A) and Focusrite Scarlet audio interface in the sound-proof booth of the Phonetics Lab at McMaster University. For all recordings, the microphone distance was specified to be around 10 cm, with the microphone being horizontally off-centre (from the lips) by ~30–45 degrees.

To ensure that the speech stimuli sounded natural without artificial manipulation or distortions, each of the voice qualities of interest was naturally (speaker-) produced. Although it can be challenging to produce several different vocal qualities on cue, we assumed that professional voice actors, as well as graduate students in Linguistics, would be highly skilled in their ability to do so. To ensure all speakers were producing exemplary productions for all vocal qualities and would sound highly natural, we explained each vocal quality, then acoustically demonstrated the vocal quality, and then continuously directed speakers on how to produce it. This included producing the vocal quality continuously throughout the produced sentences (i.e., from the start of the production to the end of the production), a comfortable and natural speech rate (not too fast, or slow), as well as limiting pitch variation (as stable and flat f0 as possible), and amplitude variation (avoiding emphasis or stress). Once speakers were able to produce each vocal quality consistently and with the previously mentioned constraints (continuous vocal quality production throughout utterance, natural speech rate, stable and flat f0, stable and consistent amplitude distribution) over the given paragraphs, they were then recorded. Both paragraphs were repeated three times for all voice qualities. The best of these repetitions was then selected as the stimulus for the listeners (i.e., the repetition with the least variation in pitch, and amplitude, continuous vocal quality production throughout the utterance, and natural speech rate). Prosodic differences were as tightly controlled as possible across speakers and conditions through the continuous direction during practice and recording sessions and auditory checks of the stimuli by the researchers. However, prosodic characteristics like f0 or intensity differences were not artificially manipulated to avoid the introduction of artifacts and did possess some variation across speakers. Since we are examining these vocal qualities against the speaker's own modal production (in other terms the baseline) we hope that any differences in prosodic control and variation across speakers' production is less impactful than if comparing directly to other speaker's productions.

The voice qualities produced were *modal*, continuous *nasalization* (specifically hypernasalization; hypernasality was produced with a lowered velum from beginning of a sentence to its end), continuous glottalization (*creaky* voice), continuous *breathy* voice, and continuous *smiling* classified into two conditions: natural (where speakers were instructed to produce a natural, comfortable smile while recording stimuli; labeled SmilingN), and extreme *smiling* (speakers were instructed to smile excessively and to an extreme while producing the stimuli; labeled SmilingEX)[10]. The smiling conditions were also visually monitored during recording sessions. Within the *breathy* voice condition, we included two distinct classes: a *natural breathy* voice production (as

[10] Note that by "continuous" we mean produced from the start to the speech production to the end of the speech production.

produced by the speaker, labeled BreathyN) and an *artificial breathy* voice production (labeled BreathyT). This *artificial breathy* production was created by taking the measurement of HNR (Harmonic to Noise Ratio) of each speaker's *natural breathy* production and overlaying a speech- frequency shaped noise signal onto their *modal* production with identical HNR measurement as the *natural breathy* production but with a rather technical (or speech transmission channel) noise overlaid[11]. In total, the stimuli consisted of 7 different vocal qualities, including a *modal voice* production for each speaker as the baseline.

After recording, the audio samples were screened with the audio editor Amadeus Pro (Hairer, 2021) and carefully checked for achieved accuracy and consistency of each vocal quality production by the two authors of this study. Additionally, a steep high pass filter (80 Hz for male speakers; 150 Hz for female speakers) was applied to remove and attenuate any additional low- frequency noise which may have been a part of the original recordings.

The final stimuli count was 84 acoustic stimuli (*six speakers* x *two paragraphs* x *seven voice quality conditions*). Stimuli were not repeated, so each acoustic stimulus was only played once for each set of questions.

In the following, we present results for measuring the acoustic parameters of the produced stimuli in the three vocal qualities *modal*, *creaky* and *breathy* to confirm that all stimuli were produced consistently and according to the specifications outlined above. The acoustic measurements used were average speech rate, average fundamental frequency, its standard deviation, CPP, HNR, jitter and shimmer. These measurements are presented in Appendix A1. Generally, we found that both male and female speakers produce the stimuli in similar and expected ways. The average speech rate is approximately four syllables per second and does not vary systematically between the modal, creaky and breathy conditions. Furthermore, the average f0 and its standard deviation within speakers also remains consistent across conditions. Speakers show the expected decreases in creaky condition (Blomgren et al., 1998; exception: speaker EM) and very similar values for modal compared to breathy voice (except for speaker HK and to some extent MK). For breathy quality, decreased values for CPP (cepstral peak prominence) measurements are an indicator of breathiness in speech with smaller ratios representing greater differences in breathiness perception (Park et al., 2019; Murton et al., 2020). We avoided using HNR measurements as errors in location of individual pitch pulse onsets can strongly affect HNR (Hillenbrand, 1987). Since the parameter CPP strongly correlates with breathiness and is more resistant to errors in fundamental period location than HNR (Hillenbrand and Houde, 1996), we examined the CPP values and found all six speakers produce consistent differences between modal and breathy voice. With respect to creaky voice, the parameters jitter and shimmer are often used as acoustic correlates. Jitter and shimmer measure the acoustic irregularities of vocal fold vibration and are linked to roughness, hoarseness or breathiness of a voice with higher measurements correlating to increases to these aspects of speech (Blomgren et al., 1998).

[11] The HNR measurement was done using Praat's algorithm using the object type "Sound: To Harmonicity: (cc)".

Specifically, jitter relates to frequency variation from cycle to cycle, while shimmer relates to amplitude variation (Murton et al., 2020). All of our speakers for almost all paragraph conditions demonstrate clear increases of jitter and shimmer in their creaky voice condition, corresponding to findings from Blomgren et al. (1998).

2.2 Participants

Twenty-seven participants took part in the perception study. They were primarily undergraduate students at McMaster University around the age of 20–23 (the majority of whom are studying Linguistics, Health Sciences or Psychology). All participants reported normal hearing and cognition. They answered a set of demographic questions including gender, age, acquired and spoken languages and musical education background. The experiment was conducted in a sound-proof booth at the Phonetics Lab at McMaster University using the Gorilla Experiment platform with wave file playback and using state-of-the-art acoustic playback conditions (Focusrite Scarlett audio interface, Sennheiser HD 598 linear frequency-response headphones). The duration of the experiment was around 60 minutes, including the pre-screening components.

2.3 Experimental setup

2.3.1. Scales

One effective way of eliciting the perception of personality traits is the use of continuous sliding scales. In voice quality and personality research, different researchers used very different types of scales. One study by Puts et al. (2007) examined the perception of dominance/authority through the use of scales to acquire ratings. The researchers posed questions to listeners about a speaker's voice, including the perception of a speaker's likelihood to win in a physical fight or the dominance or submissiveness of the speaker. Weiss and Moeller (2011) also utilized sliding scales to establish the likability of a speaker with the German antonyms *sympathisch—unsympathisch* (in English, a rough equivalent of *pleasant—unpleasant).* Several other studies (Rosenberg and Hirschberg, 2009; Berger et al., 2017; Niebuhr et al., 2018) have implemented statement- based questions, e.g., "The speaker is X," with a study-specific decision of which perceptual qualities are selected for X. Among these studies, there are variations with these statement-based questions; some are simply yes/no responses, while for others, responses are presented as a sliding scale from *strongly agree— strongly disagree*.

2.3.2. Presented statements

The statements used in the present study were based on research by Rosenberg and Hirschberg (2009) and Tskhay et al. (2018). In their research, Rosenberg and Hirschberg selected tokens based on their own judgement on whether they perceived the token as being either *charismatic* or *non- charismatic,* resulting in 26 stimuli with a mean length of 10.09s. The tokens were as context neutral as possible (e.g., "It's a pleasure to meet with you today."). For each of their 26 tokens, participants were asked directly to rate how charismatic the sample was on a 5-point scale. Additionally, participants were then asked to rate additional 23 attributes, using statements based on previous literature on charisma (see below). Examining the Big 5 in relation to the questions presented by Rosenberg and Hirschberg and the research presented by Tskhay et al., openness was not a trait applicable to ratings of charisma and was therefore omitted in the current study. Due to the high number of vocal qualities in our study's design, 10 statements were presented rather than the original 26 of the Rosenberg and Hirschberg study to prevent an excessively long experiment, and these statements were first classified by personality trait type and then balanced according to the proportion of each personality type in Rosenberg and Hirschberg's study (five extroversion, three agreeableness, one conscientiousness, and one neuroticism). Modeling after previous research (Rosenberg and Hirschberg, 2009; Tskhay et al., 2018), the statements regarding the personality traits of agreeableness, extroversion and conscientiousness were designed to have higher scores of these personality traits corresponding to higher participant ratings, meaning ratings were more positively associated with that trait. For the neuroticism personality trait, the statement aimed to have lower scores for higher participant ratings. These lower scores have been positively correlated to charisma as high scores for this trait are often associated with more negative connotations such as anxiety and proneness to negative emotions; the higher the score for neuroticism, the more the trait is exhibited, the lower the score, the less the trait is exhibited. Despite all of our statements being framed positively [rather than both positively and negatively as in Rosenberg and Hirschberg (2009)], the results should not be skewed, as scores for the Big 5 relate to either high or low scores within each trait (openness, conscientiousness, extroversion, agreeableness, and neuroticism).

For the current study, these 10 statements were presentedto the listeners. As described, these sentences were constructed by the researchers modeling the research of Rosenberg and Hirschberg (2009) and Tskhay et al. (2018) and created with a neutral valence to avoid any influence of positive or negative emotional connotations of the speech stimuli on listeners. The statements depicted the speakers as professors with the intention of establishing a relationship between the speaker and the (student) listener. Based off the previously mentioned results regarding the various vocal qualities, their uses, and the different speech environments or contexts in which they may be preferentially used, some of these vocal qualities within the current study might not be expected given the established context of an academic setting (i.e., that the speakers are "professors"). However, our rationale for labeling the speakers as professors was to prevent any other interpretations of social standing differences between different speakers, as well as between speakers and listeners, and speech environment (formal rather than colloquial like friends or family). All of these conditions could impact the perceptual ratings of speakers. Since many of the participants were university students at McMaster University, we decided it would be both interesting and relevant to characterize speakers as professors.

The ten statements were split in time, with the first screen containing the first 5 statements and the following screen containing the last five statements (Figure B1 in Appendix). This was done to prevent participants from being overwhelmed by excessive text content on one computer screen. Each screen played the audio stimuli once. Participants would slide the "button" to the desired location on each scale (for each statement) to represent how much they either agreed or disagreed with each statement presented (Figure B1 in Appendix B).

The 10 statements related to 4 of the Big 5 of personality traits:

i) extroversion (5 statements like "This professor engages students in the classroom")
ii) agreeableness (3 statements like "This professor is positive and likeable")
iii) conscientiousness (1 statement: "This professor is organized and detail oriented")
iv) neuroticism (1 statement: "This professor is convincing in the way they speak")

The listener's task was to judge the ten presented statements with respect to the simultaneously and acoustically presented audio file (and thus containing the different recorded vocal qualities). As described before, each of the audio files contained one paragraph and was recorded by the 6 individual speakers with seven different voice qualities. The statements on the screen (which were accompanied by the presented audio stimulus) would then be judged by the listeners using continuous sliding scales (from *strongly agree* to *strongly disagree)*. For our analyses, the *strongly disagree* end of the scale would be coded as 0% listener rating, and the *strongly agree* end of the scale would be coded as the 100% point of possible listener rating.

2.4 Statistical analysis

The statistical analysis was performed using the software R (R Core Team, 2018) and RStudio (R Studio Team, 2018). Parametric (e.g., ANOVA) or non-parametric (e.g., Wilcoxon) tests (depending on tests for normality of the data distributions) are used to determine whether listener responses/scores for each examined vocal quality (with respect to the presented statements and thus personality trait classification) as a dependent variable would be significantly different compared to the *modal* voice baseline (judging the exact same vocal quality stimulus and presented statement).

3 Results

3.1 Speaker-specificity vs vocal quality influences

First, we aimed to examine whether each examined speaker would indeed drive a *vocal quality difference* in participant ratings or whether participants instead chose to rate an overall and general speaker personality (i.e., a personality *gestalt*) independent of the presented vocal quality manipulations and variations. In other words, we wanted to examine if listeners indeed showed an influence of varying vocal qualities *for each speaker* or rather judged each speaker based on his/her overall *vocal personality*.

To examine this question, we first present each speaker, and each examined vocal quality for listener ratings using violin plots with overlaid boxplots, as shown in Figure 2. Firstly, it can be seen that for each speaker, the median rating for *modal voice* differs, thus establishing an overall speaker effect on listener ratings. Additionally, certain speakers are judged more positively (i.e., achieve higher response ratings overall) than other speakers, e.g., speaker CS is rated more positively overall for all examined vocal qualities than, for example, speaker JF. Furthermore, the variation of vocal quality clearly shows an effect on listener ratings, with *creaky* voice obtaining consistently negative (lower) participant ratings (except for speaker EM) and *smiling* receiving consistently positive (higher) ratings (except for speaker SA). Please note, again, that for neuroticism, the scale is inversed as higher ratings for our scale correlate to negative neuroticism personality attributes. *Breathy* and *nasal* qualities show varying results compared to *modal voice* across different speakers, but it seems that their ratings rather closely correspond to the overall *modal voice* ratings for each speaker. Finally, there does not appear to be an influence from the speaker profession on listener ratings, as can be seen in Figure 2: professional voice actors (EM, CS, SA, HK) do not show apparent rating differences compared to the Linguistics graduate students (JF, MK).

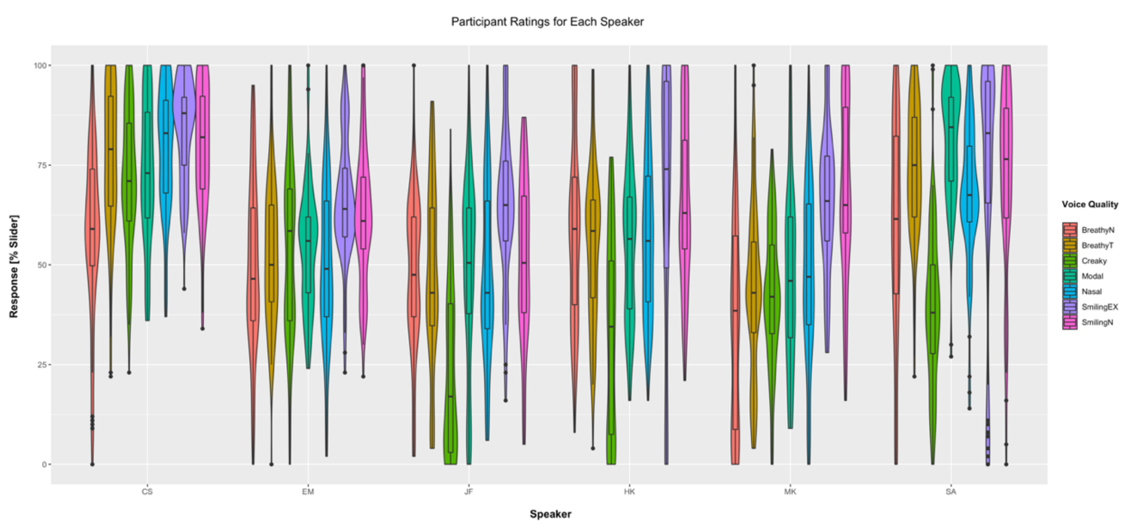


Figure 2: Voice quality ratings for all listeners, separately shown for each examined speaker (shown on the x-axis; the first three speakers are male, the last three female). The y-axis displays the slider position percentage (0% corresponding to the left extreme value of the slider and 100% to the opposite extreme). The colors represent the various examined vocal qualities. Shown are violin plots with overlaid boxplots.

Next, we present in Table 1 the correlation coefficients between *modal voice* and all other examined vocal qualities for each of the examined speakers and for all speakers combined. Theoretically, if listeners exclusively judge the acoustic personality of the underlying speaker (i.e., providing a rating of the speaker *gestalt* independent of the speaker- produced vocal quality), then correlations between modal voice and all other vocal qualities should be close to +1 (i.e., increasing *modal voice* ratings for that speaker should also increase all other vocal quality ratings simultaneously, thus excluding a possible effect of individual vocal quality on listener judgments). In contrast, if listeners exclusively judge the different vocal qualities (but choose to ignore the overall speaker identity), then correlations would strongly depend on the individual vocal quality comparisons. For example, it could be expected, based on the results in Figure 2, that the correlation between *modal* and *creaky* would be inversely related (i.e., closer to −1) compared to the correlation between *modal* and *smiling* (which could be positively correlated closer to +1), and all other comparisons showing varying correlations, but, most importantly, not being uniformly close to +1 as this would suggest an absence of a judged vocal quality difference. The correlation coefficients in Table 1 show that, for all 6 speakers, most vocal qualities obtain *varying* correlations (i.e., values *not* close to +1), thus establishing a clear influence of vocal quality on all listener ratings. For example, speaker SA shows a very high negative correlation between *modal* and *creaky* voice (i.e., if ratings for this speaker's modal increase, the ratings for creaky decrease), whereas this speaker's correlation between *modal* and *natural smiling* vocal quality condition is highly positively correlated (increased modal ratings correspond to increased *natural smiling* ratings), which clearly shows the influencing effect of *creaky* compared to *natural smiling* vocal quality on listener ratings. However, when examining the correlation table, it can also be shown that the vocal quality correlations are rather complicated and not straightforward (e.g., correlations between *creaky* and *modal* are highly positive for five speakers, and *smiling* vs. *modal* is highly negative for two speakers), but, importantly, the table, together with Figure 2, shows a clear influence of examined vocal qualities on overall listener ratings.

Speaker	Gender	Creaky vs. modal	Nasal vs. modal	BreathyN vs. modal	BreathyT vs. modal	SmilingN vs. modal	SmilingEx vs. modal
CS	Male	0.97	−0.21	−0.78	0.94	0.09	−0.79
EM	Male	0.85	0.80	0.70	0.99	−0.75	−0.65
JF	Male	1.00	0.85	0.53	0.93	0.79	0.73
MK	Female	0.72	0.32	0.29	0.96	0.26	0.40
HK	Female	0.68	1.00	0.97	0.97	1.00	0.68
SA	Female	−0.90	0.86	0.90	0.39	0.90	0.83

All speakers		0.47	0.42	0.00	0.91	0.28	0.02

Table 1: Correlation coefficients for *modal voice* vs. all other examined vocal qualities (i.e., correlating Modal-Creaky, Modal-Nasal, Modal-BreathyNatural, Modal-BreathyTechnical, Modal-SmilingNatural, Modal-SmilingExtreme), calculated separately for each speaker and combined for the six speakers.

3.2 Vocal quality influences

To examine statistical differences between the examined manipulated vocal qualities, we first conducted a Shapiro-Wilk test for normality of the data distributions for each of the vocal qualities (i.e., one test for all *modal voice* responses, one test for all *nasal voice* responses and so on). All tests for normality were highly significant (see density plots of the seven vocal qualities in Figure C1 in Appendix C; see also the distributions of each violin in the violin plot in Figure 2), so we cannot assume a normal distribution of the data and thus decided to conduct significance tests using the non-parametric Wilcoxon signed-rank (matched sample) test. We performed pairwise comparisons to determine the statistical significance (1) of each vocal quality compared to the baseline *modal voice* and furthermore (2) comparing the *natural breathy* vs. *artificially breathy* and *normal smiling* vs. *extreme smiling* vocal qualities. In sum, 8 Wilcoxon tests were conducted, and the significance values shown in Table 2 are Bonferroni-corrected for these multiple comparisons. Effect sizes comparing each examined vocal quality compared to the modal voice perception are also reported. The table shows that *all* comparisons of the six examined different vocal qualities against the *modal voice* baseline are highly significant; thus, all 6 vocal qualities obtain significantly different listener ratings when compared to the perceived *modal voice* baseline. Comparisons of the effect sizes show a medium effect size for both creaky (rated lower or more negatively compared to modal voice) and the two smiling conditions (rated higher or more positive compared to modal voice; with extreme smiling having a higher effect size). In contrast, the natural breathy condition has a small effect size (rated lower or more negatively compared to modal voice), and all other vocal qualities have negligible effect sizes. Finally, the pairwise comparison of the two *smiling* conditions shows for the Wilcoxon test that they are perceived significantly different, and the same is true for the comparison of the two *breathy* conditions, which also shows highly significant differences.

	Median, mean and standard deviation	*z*-values (Wilcoxon test)	*p*-values (Wilcoxon test)	Effect sizes (comparison to modal)
Modal	56, 53.6, 24.3	–	–	–
Creaky vs. modal	38, 38.8, 26	−26.99125	$p < 0.001$***	−0.61
Nasal vs. modal	52, 50.9, 25.6	−6.021475	$p < 0.001$***	−0.11

BreathyN vs. modal	50, 47.6, 26.1	−12.03835	$p < 0.001$***	−0.25
BreathyT vs. modal	54, 51.6, 24.2	−4.951314	$p < 0.001$***	−0.08
SmilingN vs. modal	67, 65.8, 21.6	−25.61255	$p < 0.001$***	0.50
SmilingEX vs. modal	70, 68.2, 22.2	−28.47034	$p < 0.001$***	0.60
SmilingN vs. smilingEx	-	−7.67	$p < 0.001$***	-
BreathyN vs. breathyT	-	−8.37	$p < 0.001$***	-

Table 2: Mean, median and standard deviations for each vocal quality (the bold-printed vocal quality values of the first column are reported) and results of the statistical *Wilcoxon signed-rank pairwise significance test* (z-values: third column; *p*-values: fourth column) comparing the two vocal qualities stated in column one, calculated over all participant responses. All p-values are Bonferroni-corrected. The last column gives the effect sizes (Cohen's d) for each examined vocal quality compared to modal voice perception. The *** symbol indicates statistically significant at the 0.1% level ($p < 0.001$).

3.3 Personality traits vs vocal quality influences

In the following, we aim to examine the interaction between *vocal quality variation* and perceived *personality traits*. Figure 3 shows violin plots with overlaid box plots over all speakers, split by the four examined personality traits (x-axis) and examined vocal quality (colors). The baseline would be the rating of the perceived *modal voice*, and it can be seen that this vocal quality shows very similar values when comparing the four personality traits. Visual examination of the vocal quality differences for each personality trait confirms the results of the previously presented significance tests: mean listener ratings were higher, or more positive, for *smiling* for both the *natural smiling* condition and the *extreme smiling* condition across all personality traits. Again, the *extreme smiling* condition is rated higher compared to the *natural smiling* condition, and the natural breathy condition is rated lower than the artificial breathy one, corresponding to the significant differences observed in Table 2. Inversely, *creaky* was perceived lower, or more negatively, for all four personality traits. This is in line with previous research (Tartter, 1980) that general perceptions of *smiling* are correlated to more positive emotions and associations like trustworthiness, friendliness, etc., while *creaky* is perceived more negatively.

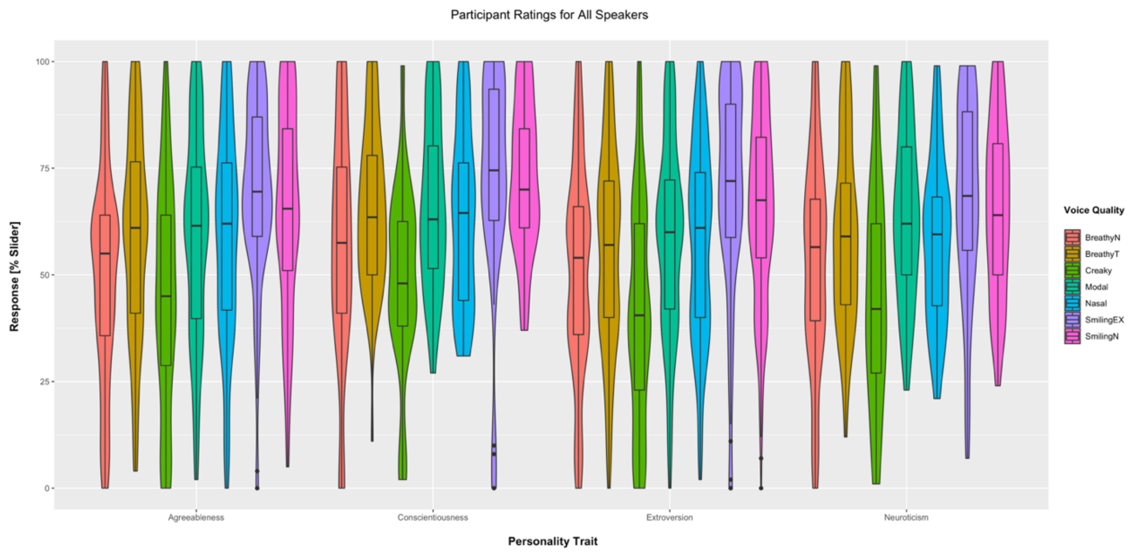


Figure 3: Voice quality ratings for all listeners aggregated over all six speakers' productions. The y-axis displays the slider position percentage (0% corresponding to the left extreme value of the slider and 100% to the opposite extreme). The x-axis shows the aggregation of the 10 statements into the four personality traits of interest: agreeableness, conscientiousness, extroversion, and neuroticism. The colors represent the various examined vocal qualities. Shown are violin plots with overlaid boxplots.

While Figures 2, 3 show the differences between the seven examined vocal qualities and thus gave an appropriate first overview of the obtained listener responses and their response distributions, the main aim of this study is to investigate the difference between an observed vocal quality and its corresponding *modal voice* perception, or, in other words, to see the pure effect of each vocal quality manipulation with respect to the four personality traits. In order to see this effect, we calculated, for each vocal quality judgement, the *difference percentage* between examined vocal quality and the baseline *modal voice* for that exact same acoustic stimulus comparison, thus effectively providing a pure effect of each vocal quality on listener ratings, split by personality trait. For example, we took the judgement of listener 1 judging the first paragraph of speaker 1 produced in *modal* voice and subtracted this value from the judgement of listener 1 judging the first paragraph of speaker 1 in a *creaky* voice, thus providing a measurement value showing the absolute difference in vocal quality rating (compared to *modal voice* judgements) for that specific speaker, listener, and paragraph identity. This calculation was then performed for all other (vocal quality, speaker, and listener) judgements. Thus, this difference quantifies the effect of the magnitude of change in vocal quality without taking into account other parameters. The results are presented in Figure 4, again as violin plots with overlaid boxplots. As can be seen, *smiling* again has the most considerable influence on all personality trait ratings, with more pronounced effects on agreeableness, conscientiousness and extroversion and a much smaller effect on neuroticism. In contrast, the *creaky* vocal quality has the strongest negative effect, with the most significant effect shown for conscientiousness compared to the other three traits. *Breathy voice* has a smaller negative effect on listener ratings, and interestingly this negative effect is strongest for neuroticism. An interesting result is the comparison of the two within-categories: the *natural smiling* vs. *extreme smiling,* and the *natural breathy* vs. *artificial breathy* condition. We do not observe the expected ceiling effect for *extreme smiling* conditions. Instead, for three of the four personality traits (excluding neuroticism), the *extreme smiling* condition consistently outperforms the *natural smiling* one, thus increasing the positive listener rating for more extreme smiling of each examined speaker. Interestingly, the *artificial breathy* condition generates more positive listener ratings compared to the *natural breathy* condition (see also Figures 2, 3 overall ratings), or, to turn it around, the *natural breathy* condition consistently leads to lower, or more negative, listener ratings compared to *artificial breathy* productions.

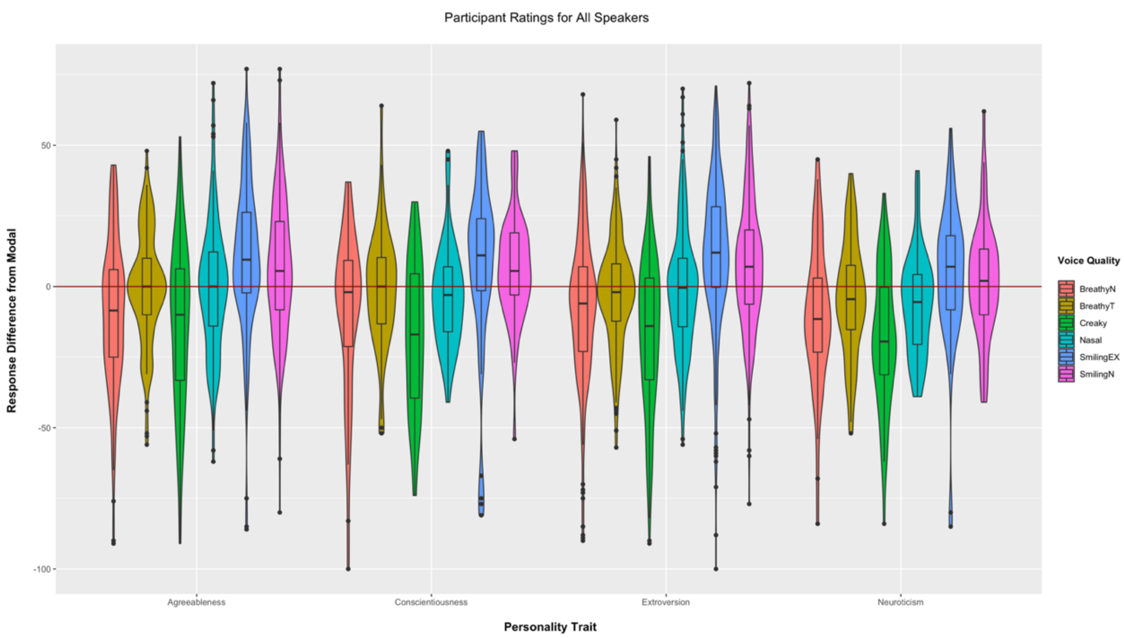


Figure 4: Violin plot and overlaid boxplots of the point-wise difference ratings for all listeners aggregated over all 6 speakers' productions. The y-axis displays the percentage difference as compared to each baseline (modal voice), i.e., the difference between each vocal quality rating and the corresponding modal voice quality (see text for further explanation). Positive rating difference percentages (compared to modal voice ratings) are above the 0% line; negative rating percentages are below this line. See Figure 3 for a description of axes and colors.

3.4 Effects of *speaker* and *listener* gender

When examining the effect of *speaker* gender, Figure 5 shows the mean differences in listener ratings, comparing listener judgements separately for male and female *speakers*, the produced vocal qualities and the four personality traits. We also provide results of the Wilcoxon pairwise significance test in Table 3, over all speakers and for each examined vocal quality. Overall, the gender of the *speaker* has a significant effect on listener ratings ($p < 0.0001$), and all vocal qualities except *natural breathy* voice show a highly significant effect of speaker gender (see Table 3). Examination of the means in Figure 5 shows that there is a tendency that female speakers are judged more positively, independent of the examined personality trait and for almost all vocal qualities. Furthermore, we consistently see larger differences for two vocal qualities: *creaky* and, to some extent, *smiling*. Listeners rated female speakers more negatively when producing *creaky* voice compared to males for all personality traits. Our results thus confirm previous research (Anderson et al., 2014; Chao and Bursten, 2021) that demonstrated that *creaky* voice is frequently perceived negatively in women in a variety of environments. Additionally, both *smiling* variants are consistently rated higher for female speakers than their male counterparts.

	Speaker gender		Listener gender	
	z-score	*p*-value	*z*-score	*p*-value
All (qualities)	−7.5408	<0.001***	−4.3274	<0.001***
Modal	−6.9697	<0.001***	−2.4725	0.01342
Creaky	−13.336	<0.001***	−2.1832	0.2367
Nasal	−4.976	<0.001***	−2.3980	0.01648
BreathyN	−1.1722	0.2391	−0.1885	0.8504
BreathyT	−5.4873	<0.001***	−3.74462	<0.001***
SmilingN	−14.0154	<0.001***	−3.40800	<0.001***
SmilingEx	−14.0154	<0.001***	−3.40800	0.2337

Table 3 Results of the *Wilcoxon signed-rank pairwise significance test* (*z*-values and *p*-values) for each examined vocal quality and aggregated over all vocal qualities. All p-values are Bonferroni-corrected (due to multiple comparisons). The *** symbol indicates statistically significant at the 0.1% level ($p < 0.001$).

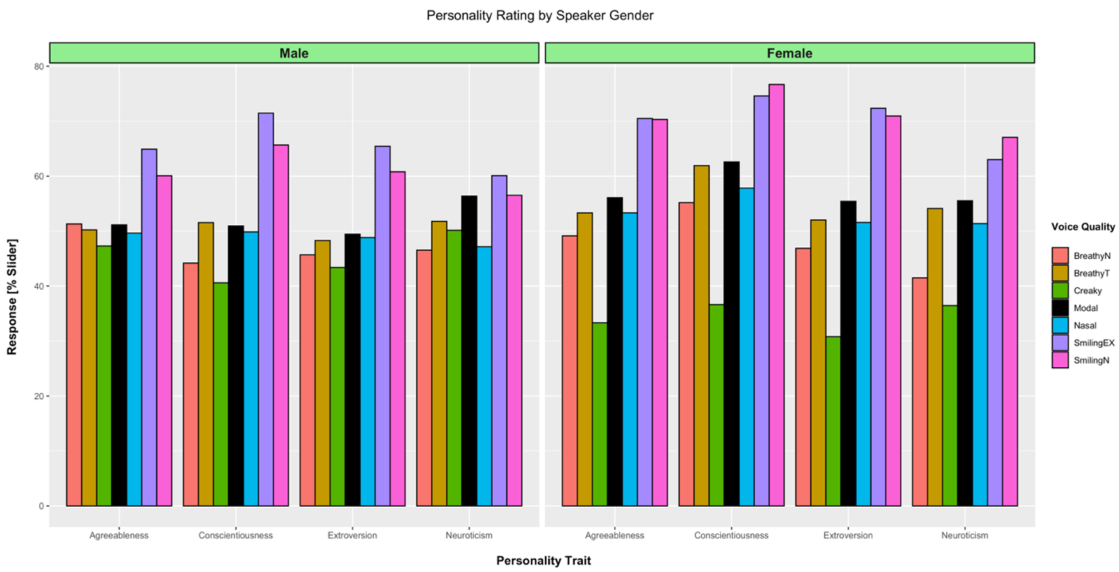


Figure 5: Mean ratings over all listeners comparing male *speakers* (left) versus female *speakers* (right) productions, split by vocal quality. See figure 3 for a description of axes and colours.

Figure 6 shows mean plots comparing the effects of male and female *listener* gender on personality ratings, split by vocal quality and personality trait. Table 3 provides the significance results for each vocal quality. Overall, *listener* gender, similar to *speaker* gender, also has a significant effect on listener ratings ($p < 0.001$), however, the only vocal qualities rated significantly different when comparing the two (listener) genders are the *natural smiling* ($p < 0.001$) and *artificial breathy* ($p < 0.001$) vocal quality.

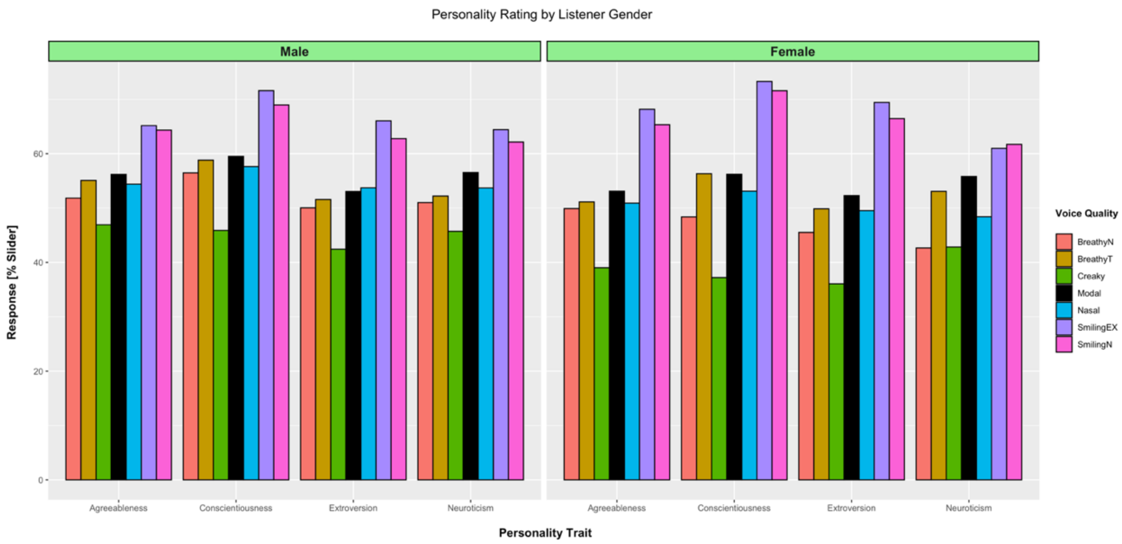


Figure 6: Mean ratings over all listeners comparing male *listeners* (left) versus female *listeners* (right) productions split by vocal quality. See figure 3 for a description of the axes and colours.

4 Discussion

Our results demonstrate that variation of the four different vocal qualities *breathy*, *creaky*, *nasal* and *smiling,* varied for the same individual speaker, can strongly and significantly influence listener perception of that speaker's personality traits, both positively and negatively: Results of the conducted Wilcoxon significance tests (see Table 2) show significantly higher listener rating scores for *smiling* voice qualities (for both *natural smiling* and *extreme smiling* condition) across all examined personality traits, whereas the *creaky* vocal quality was consistently and significantly rated lower for all personality traits for all participants, and thus perceived more negatively overall. For H1 we found that the continuous production of this creaky voice negatively impacts ratings of all personality traits. These results thus confirm the results of previous research regarding *creaky* voice and its unfavorable perception by listeners, which has linked this vocal quality to impressions of boredom and sadness (Gobl and Ní Chasaide, 2003), which would be classified into low neuroticism scores[12]. When specifically looking at *creaky* voice produced by women, the production of this vocal quality can negatively affect the ratings of competence (i.e., lower score in conscientiousness trait), trustworthiness (i.e., lower score in conscientiousness), and education level (Anderson et al., 2014). Furthermore, our results show this vocal quality is more negatively rated as compared to other vocal qualities such as *breathy* or *nasal* voice, both currently with very limited research results. Our results however are in contrast to those of both Yuasa (2010) and Pittam (1987), which found *creaky* voice correlated positively with professionalism, intelligence, friendliness, genuineness, and nonaggressiveness, as well as positive assumptions about a speaker, like assumed higher level of education. In

[12] Based on our inverted scale for neuroticism described in the Methods section.

sum, our study's lower ratings for all personality traits in combination with the results of previous studies (showing *decreased* scores for neuroticism, and conscientiousness being indicative of lack of charisma) all suggest that (continuously produced) creaky voice decreases the perception of speaker charisma. We can therefore accept this hypothesis.

For H2 our results are consistently and significantly higher ratings (see Table 2) for smiling. This confirms previous research, which found that *smiling* in women signals trustworthiness (high score in conscientiousness), indicated warmth (high score in extroversion trait) and enthusiasm (high score in extroversion trait) to the listener, whereas men who were *smiling* were interpreted as lacking self-doubt (high score in conscientiousness), confidence (high score in conscientiousness), and calmness (high score in neuroticism[9]; Vazire et al., 2009). This adds to general perceptions of *smiling* which are correlated to more positive emotions and associations like trustworthiness and friendliness (high score in extroversion; Tartter, 1980). Therefore, for H2 we can accept this hypothesis. Inversely to creaky voice, for *smiling* the observed higher ratings for all examined personality traits, in combination with the results of previous studies (showing *higher* ratings for neuroticism, extroversion, and conscientiousness being indicative of charisma), suggest that *smiling* positively impacts perceptions of speaker charisma.

For H3, also examining smiling, we find interesting results contrary to those presented by Tschinse et al. (2022). Our results in Table 2 show statistically significant differences between SmilingN (natural) and SmilingEX (extreme), with SmilingEX outperforming SmilingN with respect to positive listener ratings. These results thus reject our H3 null hypothesis. Although visually the differences between SmilingN (natural) and SmilingEX (extreme) appear rather small (see Figure 5), our statistical analysis demonstrates that increasing the smiling dimension also increases the positive influences on personality traits perception and therefore charisma. Whether the differences in results comparing our data with Tschinse et al. (2022) are due to our within-subject design or rather other methodological differences remains a cause for further study. Some of these mentioned methodological differences could be a result of stimuli: the stimuli in our study used short, isolated paragraphs (approximately 12 seconds) while the stimuli of Tschinse et al. (2022) were longer 1-min pitches. Prolonged auditory stimuli input allows for more habituation for participants and "saturation". Furthermore, the instruction for our extreme smiling and permanent smiling are not exclusively interchangeable, with the latter being temporally defined while the former is not.

For the other two vocal qualities, *nasal* and *breathy,* no consistent and robust differences in listener ratings across speakers could be found. However, the Wilcoxon significance tests showed that these two qualities still obtained statistically significant differences, all compared to the *modal voice* baseline (see Table 2). Despite this fact, our results for both *breathiness* and *nasality* do not suggest a strong and clear trend relating these voice qualities to individual perceived personality trait differences since overall ratings of *nasality* and *breathiness* follow very similar trends as the *modal voice* baseline.

Our results thus suggest that both nasality and breathiness do not play a salient role compared to smiling or creaky voice for personality trait attribution, although they both seem to lower listener scores for several traits for most speakers. For *nasality*, this result is quite interesting as it fills a current gap in the literature regarding the saliency and influence of *nasality* on personality trait perception and charisma. For *breathy* voice, previous research has suggested that this vocal quality influences perceptions of a speaker's perceived sexuality and sensuality, but only when the speaker is female (Laver, 1980). Also, solidarity perception with speakers is higher for *breathy* voices (Pittam, 1985). From our data, the results suggest that *breathiness* does not have that strong influence (but again, please note the significant difference to modal voice based on the Wilcoxon test).

Together, the results show increases or decreases in listener ratings for each vocal quality type. They appear to either *all* increase or decrease together, depending on the positive or negative perception of that vocal quality. By interpreting these traits collectively, we can see that those general increases/decreases of personality trait perceptions have a relationship with charisma; the higher, more positive, the ratings of traits, the greater the perception and saliency of charisma, whereas the lower, more negative the ratings, the lower the perception and saliency of charisma. Since we found that these increases/decreases in ratings synchronize across the different personality traits of the Big 5 (within each vocal quality), this can aid in future research on charisma in two ways: On the one hand, not all personality traits (of the Big 5) need to be utilized in experimental designs (i.e., only using questions/statements framed within the Extroversion trait, or Agreeableness, etc.) in order to capture meaningful interpretations of charisma and its presence. On the other hand, although statistically significant, some vocal qualities (nasal and breathy) are less salient in charismatic perception than other vocal qualities (creaky voice and smiling).

Further investigation into our within-category breathy voice differences reveal a statistically significant difference between BreathyN (natural breathy condition) and BreathyT (technical breathy condition). BreathyT is perceived with higher personality trait ratings and thus charisma, or, turning it around, BreathyN is perceived worse, thus confirming our H5. Despite being less salient than creaky vocal quality for charismatic trait attribution, these results suggest there is indeed a difference between adding the same type of noise (speech-shaped noise with identical HNR) to either the speaker's laryngeal signal (i.e., natural speaker-produced) or to the general communication channel (thus not being modulated by speech production differences). We speculate here that listeners are indeed able to separate the added channel noise from the speaker (personality trait) judgments, thus pointing to a hypothesis that added channel noise is not as detrimental to personality trait perception as noise directly produced by the speaker's larynx.

With respect to perceptual saliency and the magnitude of participant rating differences, certain vocal qualities are more pronounced than others for personality attribution. For example, see the difference in point-by-point comparison of *creaky* vocal quality (compared to *modal*) vs. *nasal* vocal quality (also compared to *modal*) as shown in Figure 4. Although previous literature has shown the various impacts of how speakers are judged by listeners regarding vocal quality differences, the current results can provide a better understanding which vocal qualities may require more focus when attempting to increase charisma perception: *smiling* more and avoiding continuous *creaky* voice appear to be more relevant and more salient than avoiding *nasal* or *breathy* productions.

Furthermore, we examined the effect of speaker- and listener-dependent factors, namely the effect of gender for both *speaker* and *listener*. For H4, rating differences comparing male and female *speakers* and their vocal quality show consistent differences for both *creaky* and *smiling* vocal quality but not to the same extent for the other vocal qualities. Listeners rated female speakers more negatively in *creaky* voice than the corresponding male speakers for the personality traits of agreeableness, extroversion, and neuroticism. Here, our results confirm previous research (Anderson et al., 2014; Chao and Bursten, 2021) that demonstrated that *creaky* voice is frequently perceived negatively in women in a variety of environments. Additionally, both *smiling* variants are consistently rated higher for female speakers than their male counterparts. In sum, our results confirm that gender strongly influences the perception of vocal quality, both overall and within different personality trait contexts and we can accept this hypthothesis.

One limitation of the present study is the relatively small number of speakers and the observed inter-speaker variation (see, e.g., Figure 2, where listener ratings for speaker EM show the opposite pattern for *creaky* vs. *modal* voice compared to the other five speakers). Six speakers can provide a general picture of vocal quality and personality attribution, but this picture is still limited in the scope of potential variation, which may naturally occur in the production of individual speech patterns. Also, as can be seen in the correlation table, vocal qualities across speakers are not judged uniformly, thus introducing speaker- specific variation in this vocal quality study. For future studies, a higher number of different speakers could provide a more detailed understanding of the effect of vocal quality variations on charismatic traits, and it might continue to examine the more fine-tuned effects of *nasality* and *breathy* voice (in both technical and natural variation) that gave significant overall differences compared to *modal voice* perceptions but failed to provide a clear trend of the effects on individual personality traits.

An additional point of limitation in the current study is that we could not control for several other factors of variability for perceived vocal quality and personality ratings. For example, different possible settings would feed into the concept of *charismatic* speech and influence the ratings, for example: different communication contexts (formal vs. informal), environmental settings (e.g., academic, as in this study, vs. peer ratings), types of audiences (e.g., interviewers vs. colleagues), and of course whether the purpose of the communication is to be persuasive. Specifically, the established speaker-listener relationship in our study (i.e., the speaker being defined as a professor for our student participants) could influence vocal qualities to be perceived differently than if that relationship would have been established with a different social relationship paradigm (e.g., the listener is not a student, the listener is rating a friend, the listener is rating a co-worker, etc.). Since we chose this university setting—as previously explained— there is of course the chance that some of these voice qualities could be perceived differently for different speaker settings (friends or family) or different social environments (e.g., work, socializing with friends).

Finally, it is essential to note that other factors in combination with vocal quality appear to play a role, such as age vs. creakiness (Esling, 1978; Scherer, 1979) or the interaction of creakiness, f0, and speech rate (Parker and Borrie, 2018). These, along with many other variations, suggest that, of course, vocal quality is not the only important component for listeners when giving ratings of personality traits. For future studies, the inclusion of other speech features like f0 variation, speech rate differences etc., could provide a further understanding of the interactions between linguistic speech variation and voice quality.

5 Ethics Statement

The studies involving human participants were reviewed and approved by McMaster Research Ethics Board. The patients/participants provided their written informed consent to participate in this study.

6 Author Contributions

SP and DP designed the experiment and stimuli, analyzed and interpreted data, and drafted and developed the manuscript. SP recorded the stimuli and collected participant data. DP also supervised the experiment. Both authors contributed to the article and approved the submitted version.

7 Funding

This work was supported by the NSERC Discovery Grant PGGPIN-2018-06518 to the second author.

8 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

9 References

Abdelli-Beruh, N.B., Wolk, L. and Slavin, D., 2014. Prevalence of Vocal Fry in Young Adult Male American English Speakers. *Journal of Voice*, 28(2), pp.185-190.

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Anderson, R., Klofstad, C., Mayew, W. and Venkatachalam, M., 2014. Vocal Fry May Undermine the Success of Young Women in the Labor Market. *PLoS ONE*, 9(5), p.e97506.

Antonakis, J., Bastardoz, N. & Jacquart, P. 2016. Charisma: an ill-defined and ill-measured gift. *Annual Review of Organizational Psychology and Organization Behaviour,* 3, pp.293-319.

Apple, W., Streeter, L. and Krauss, R., 1979. Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37(5), pp.715-727.

Berger, S., Niebuhr, O. and Peters, B., 2017, March. Winning over an audience–A perception-based analysis of prosodic features of charismatic speech. In *Proc. 43rd Annual Conference of the German Acoustical Society, Kiel, Germany*, pp.1454-1457.

Blomgren, M., Chen, Y., Ng, M. L., & Gilbert, H. R. 1998. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *The Journal of the Acoustical Society of America*, *103*(5), pp.2649-2658.

Bono, J. E., & Judge, T. A. 2004. Personality and transformational and transactional leadership: a meta-analysis. *Journal of applied psychology*, *89*(5), pp.901.

Borkowska, B. and Pawlowski, B., 2011. Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, *82*(1), pp.55-59.

Chao, M. and Bursten, J.R., 2021. Girl Talk: Understanding Negative Reactions to Female Vocal Fry. *Hypatia*, *36*(1), pp.42-59.

Cheng, J.T., Tracy, J.L., Ho, S. and Henrich, J., 2016. Listen, follow me: Dynamic vocal signals of dominance predict emergent social rank in humans. *Journal of Experimental Psychology: General*, *145*(5), pp.536.

Collins, S.A., 2000. Men's voices and women's choices. *Animal behaviour*, *60*(6), pp.773-780.

Costa, P. T. & McCrae, R.R. (1992). *NEO PI-R Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Esling, J. H., 1978, Voice Quality in Edinburgh: A Sociolinguistic and Phonetic Study. *Ph.D. Dissertation.* The University of Edinburgh (United Kingdom).

Esling, J., Moisik, S., Benner, A., & Crevier-Buchman, L., 2019. Voice Quality The Laryngeal Articulator Model (No. 1). Cambridge University Press.

Gick, B., Wilson, I. and Derrick, D., 2013. 'Articulating Laryngeal Sounds' in *Articulatory phonetics*. John Wiley & Sons, pp.109.

Gobl, C. and Ní Chasaide, A.N., 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, *40*(1-2), pp.189-212.

Gray, R., 2017, Boundless psychology. *Lumen*. Available at: https://courses.lumenlearning.com/boundless-psychology/chapter/trait-perspectives-on-personality/ [Accessed March 30, 2022].

Hairer, M. (2021) [software]. Amadeus Pro (version 2.8.8). London, United Kingdom; HairerSoft.

Hanson, H.M., 1997. Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America*, *101*(1), pp.466-481.

Hart, S.D. and Hare, R.D., 1994. Psychopathy and the Big 5: Correlations between observers' ratings of normal and pathological personality. *Journal of Personality Disorders*, *8*(1), pp.32.

Hillenbrand, J. 1987. A methodological study of perturbation and additive noise in synthetically generated voice signals. *Journal of Speech, Language, and Hearing Research*, *30*(4), pp.448-461.

Hillenbrand, J., & Houde, R. A. 1996. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research*, *39*(2), pp.311-321.

Hodges-Simeon, C.R., Gaulin, S.J. and Puts, D.A., 2010. Different vocal parameters predict perceptions of dominance and attractiveness. *Human Nature*, *21*(4), pp.406-427.

Imhof, M., 2010. Listening to voices and judging people. *The Intl. Journal of Listening*, *24*(1), pp.19-33.

John, O.P. and Srivastava, S., 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, *2*(1999), pp.102-138.

Klofstad, C.A., Anderson, R.C. and Peters, S., 2012. Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1738), pp.2698-2704.

Ladefoged, P.N. & Johnson, K., 2011. Airstream Mechanisms and Phonation Types. In *A course in phonetics*. Boston etc.: Wadsworth.

Laver, J., 1972. Voice quality and indexical information readings, in: *Communication in face to face interaction,* Laver, J. & Hutcheson, S. (eds.), Penguin, pp.189-203.

Laver, J., 1980. The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, *31*, pp.1-186.

McCrae, R.R. and John, O.P., 1992. An introduction to the five‐factor model and its applications. *Journal of personality*, *60*(2), pp.175-215.

Michalsky, J., & Niebuhr, O. 2019. Myth busted? Challenging what we think we know about charismatic speech. *AUC PHILOLOGICA, 2019(2),* pp.27-56.

Möbius, B., 2003. Gestalt Psychology meets phonetics–An early experimental study of intrinsic F0 and intensity. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pp. 2677-2680.

Murton, O., Hillman, R., & Mehta, D. 2020. Cepstral peak prominence values for clinical voice evaluation. *American Journal of Speech-Language Pathology*, *29*(3), pp.1596-1607.

Niebuhr, O., Skarnitzl, R. and Tylečková, L., 2018, June. The acoustic fingerprint of a charismatic voice-Initial evidence from correlations between long-term spectral features and listener ratings. In *Proceedings of Speech Prosody*, 9, pp.359-363.
Niebuhr, O. and Fischer, K., 2019, September. Do not Hesitate!-Unless You Do it Shortly or Nasally: How the Phonetics of Filled Pauses Determine Their Subjective Frequency and Perceived Speaker Performance. In *INTERSPEECH*, pp.544-548.
Norman, W.T., 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The journal of abnormal and social psychology*, *66*(6), pp.574.
Novák-Tót, E., Niebuhr, O. and Chen, A., 2017, August. A gender bias in the acoustic-melodic features of charismatic speech?. In *INTERSPEECH*, pp.2248-2252.
Park, Y., Perkell, J.S., Matthies, M.L. and Stepp, C.E., 2019. Categorization in the perception of breathy voice quality and its relation to voice production in healthy speakers. Journal of Speech, Language, and Hearing Research, 62(10), pp.3655-3666.
Parker, M.A. and Borrie, S.A., 2018. Judgments of intelligence and likability of young adult female speakers of American English: The influence of vocal fry and the surrounding acoustic-prosodic context. *Journal of Voice*, *32*(5), pp.538-545.
Pittam, J.,1985. Voice quality: its measurement and functional classification.
Pittam, J., 1987. The long-term spectral measurement of voice quality as a social and personality marker: a review. *Language and Speech*, *30*(1), pp.1-12.
Pointer, N.F., van Mersbergen, M. and Nanjundeswaran, C.D., 2022. Listeners' Attitudes Towards Young Women With Glottal Fry. *Journal of Voice*.
Puts, D.A., Hodges, C.R., Cárdenas, R.A. and Gaulin, S.J., 2007. Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), pp.340-344.
Quené, H., Boomsma, G. and van Erning, R., 2016. Attractiveness of male speakers: effects of voice pitch and of speech tempo. In *Proceedings of Speech Prosody*, *8*, pp.1086-1089.
RStudio Team. 2009-2018. *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.
R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Reetz, H. and Jongman, A., 2020. *Phonetics: Transcription, production, acoustics, and perception*. John Wiley & Sons.
Rosenberg, A. and Hirschberg, J., 2009. Charisma perception from text and speech. *Speech Communication*, *51*(7), pp.640-655.
Scherer, K.R., 1979. Voice and speech correlates of perceived social influence in simulated juries. *Language and social psychology*, pp.88-120.
Shor, R.E., 1978. The production and judgment of smile magnitude. *The Journal of General Psychology*, *98*(1), pp.79-96.

Shue, Y.-L. 2010, The voice source in speech production: Data, analysis and models. UCLA dissertation. [version 1.37 - Jun 2, 2020]
Signorello, R., & Demolin, D., 2013,. The physiological use of the charismatic voice in Political speech, *INTERSPEECH*, pp.987-991.
Tartter, V.C., 1980. Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & psychophysics*, *27*(1), pp.24-27.
Teixeira, J. P., Oliveira, C., & Lopes, C. 2013. Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, *9*, pp.1112-1122.
Torre, I., 2014. Production and perception of smiling voice. In *Proceedings of the First Postgraduate and Academic Researchers in Linguistics at York (PARLAY 2013) Conference. York, UK*.
Tsantani, M.S., Belin, P., Paterson, H.M. and McAleer, P., 2016. Low vocal pitch preference drives first impressions irrespective of context in male voices but not in female voices. *Perception*, *45*(8), pp.946-963.
Tschinse, L. M., Asadi, A., Gutnyk, A., & Niebuhr, O. 2022. Keep on smiling...? On the sex-specific connections between smiling duration and perceived speaker attributes in business pitches. In Proc. 11th International Conference of Speech Prosody, pp.1-5.
Tskhay, K. O., Zhu, R., Zou, C., & Rule, N. O., 2018. Charisma in everyday life: Conceptualization and validation of the General Charisma Inventory. *Journal of Personality and Social Psychology*, 114(1), pp. 131–152.
Tull, R.G., 1999. *Acoustic analysis of cold-speech: Implications for speaker recognition technology and the common cold*. Northwestern University.
Vazire, S., Naumann, L.P., Rentfrow, P.J. and Gosling, S.D., 2009. Smiling reflects different emotions in men and women. *Behavioral and Brain Sciences*, *32*(5), pp.403-405.
Vergauwe, J., Wille, B., Hofmans, J., & De Fruyt, F. 2017. Development of a Five-Factor Model charisma compound and its relations to career outcomes. *Journal of Vocational Behavior*, *99*, pp.24-39.
Warren, D.W., Hairfield, W.M., Seaton, D., Morr, K.E. and Smith, L.R., 1988. The relationship between nasal airway size and nasal-oral breathing. *American Journal of Orthodontics and Dentofacial Orthopedics*, *93*(4), pp.289-293.
Watterson, T. and Emanuel, F., 1981. Observed effects of velopharyngeal orifice size on vowel identification and vowel nasality. *The Cleft Palate Journal*, *18*(4), pp.271-278.
Wiener, H.J. and Chartrand, T.L., 2014. The effect of voice quality on ad efficacy. *Psychology & Marketing*, *31*(7), pp.509-517.
Wolf, N., 2015. Young women, give up the vocal fry and reclaim your strong female voice. *The Guardian*, *24*.
Yuasa, I.P., 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women?, *American Speech*, *85*(3), pp.315-337.
Ziegler, M., Danay, E., Schölmerich, F. and Bühner, M., 2010. Predicting academic success with the Big 5 rated from different points of view: Self-rated, other rated and faked. *European Journal of Personality*, *24*(4), pp.341-355.

Appendices

Appendix A

Acoustic measurements of stimuli

TABLE A.1: *The results of acoustic stimuli analysis separated by male speakers (CS, EM, JF) and female speakers (HK, MK, SA). Measurements are given for each speaker, examined voice quality (Modal, Creaky, Breathy), and paragraph identity (either 1 or 2) and were measured using VoiceSauce (Shue, 2010). The measured acoustic parameters are: speech rate (number of syllables per second), mean fundamental frequency and standard deviation (in brackets) in Hz, cepstral peak prominence (CPP; in dB), jitter (in percent), shimmer (in percent), and harmonics-to-noise ratio (HNR; in dB). For HNR, the measurements are from HNR05, within VoiceSauce, measures the HNR between 0-500 Hz. For Breathy columns, first value is BreathyN and the value in brackets is BreathyT.*

MALE SPEAKERS		**CS**			**EM**			**JF**		
	Paragraph	Modal	Creaky	Breathy	Modal	Creaky	Breathy	Modal	Creaky	Breathy
Speech Rate (syll/s)	1	4.5	4	3.6	4.2	4.1	4.3	5.3	4.7	4.7
	2	4.3	3.7	3.7	4	3.9	3.6	5.1	4.1	4.3
Mean f0 (Hz)	1	107 (21.4)	84 (94.5)	101 (26.2)	94 (13)	100 (25)	86 (9.6)	128 (31)	120 (6)	132 (18.6)
	2	108 (16.6)	85 (74)	94 (26.2)	95 (28)	109 (54)	89 (6.2)	121 (6)	122 (14)	132 (9.3)
CPP (dB)	1	19	19	18 (18)	18	17	17 (17)	17	16	16 (16)
	2	19	19	18 (18)	19	17	16 (18)	17	16	15 (16)
Jitter (%)	1	2.92	3.25	3.35	2.22	2.79	2.47	2.08	6.37	2.38
	2	2.57	3.63	2.64	2.23	3.59	2.1	1.95	6.69	2.6
Shimmer (%)	1	12.7	14.2	11.5	11.3	11.4	10.9	12.2	19.2	12.1
	2	15.1	13.6	13.2	11.9	14.5	11.4	11.3	16.9	12.1
HNR (dB)	1	22	16	22 (22)	20	15	18 (19)	17	11	17 (18)
	2	23	17	20 (22)	21	16	17 (19)	16	10	16 (17)

FEMALE SPEAKERS		**HK**			**MK**			**SA**		
	Paragraph	Modal	Creaky	Breathy	Modal	Creaky	Breathy	Modal	Creaky	Breathy
Speech Rate (syll/s)	1	5	4.5	4.2	4.2	3.9	3.8	4.6	4.1	4.2
	2	3.9	3.5	3.8	4.8	4.5	4.1	4.1	3.8	3.9
Mean f0 (Hz)	1	160 (47)	140 (57)	204 (35)	174 (19)	168 (37.6)	191 (29)	189 (37.3)	166 (27.7)	190 (32)
	2	161 (44)	154 (42)	197 (28)	173 (27)	159 (41.9)	188 (19)	196 (41)	158 (24.1)	194 (41)
CPP (dB)	1	16	18	15 (16)	19	17	17 (18)	16	17	15 (16)
	2	16	18	16 (16)	19	17	17 (18)	17	17	15 (16)
Jitter (%)	1	2.15	3.19	2.14	2.12	3.26	1.47	3.01	4.14	2.42
	2	2.38	2.5	2.13	1.71	3.51	1.62	2.1	2.44	2.04
Shimmer (%)	1	10.5	14.3	11.3	10.3	14.8	9	9.1	10.8	8.6
	2	8.6	7.9	7.6	10.2	14.3	8.6	9	8.6	8.2
HNR (dB)	1	20	14	27 (21)	27	20	30 (28)	35	32	32 (30)
	2	24	22	27 (24)	28	19	30 (28)	36	33	33 (30)

Statements

The statements used in this study were:
Extroversion Questions (5)
  Question 1: This professor engages students in the classroom
  Question 2: This professor has a strong presence in a lecture hall or on Zoom
  Question 3: This professor knows how to lead a group
  Question 4: This professor is enthusiastic about teaching
  Question 5: This professor is captivating when speaking
Agreeableness Questions (3)
  Question 1: This professor can get along with any student
  Question 2: This professor makes students feel comfortable
  Question 3: This professor is positive and likeable
Conscientiousness Question (1)
  Question 1: This professor is organized and detail oriented.
Neuroticism Question (1)
  Question 1: This professor is convincing in the way they speak

Appendix B

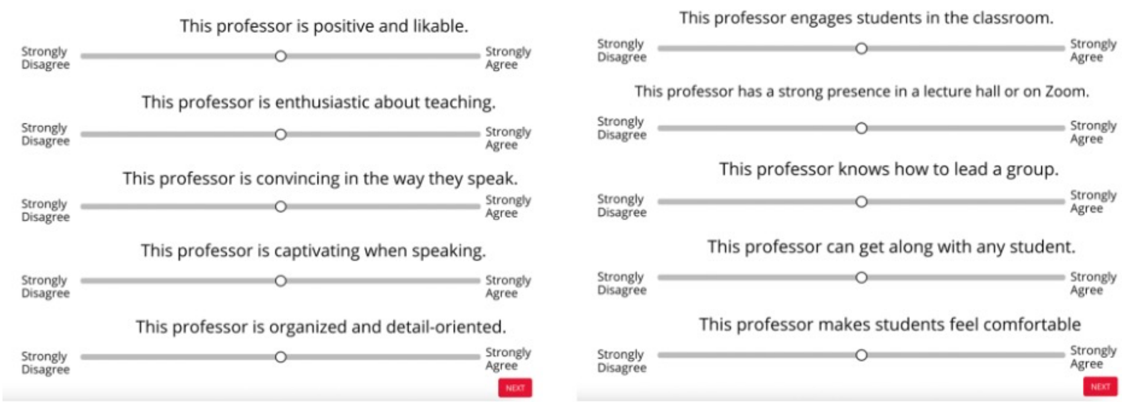


Figure B.1: Screenshot of the experimental setup with the questions presented to the listeners during the perception study shown is the screen showing the first five questions (left panel) and the second screen showing the last 5 questions (right panel).

Appendix C

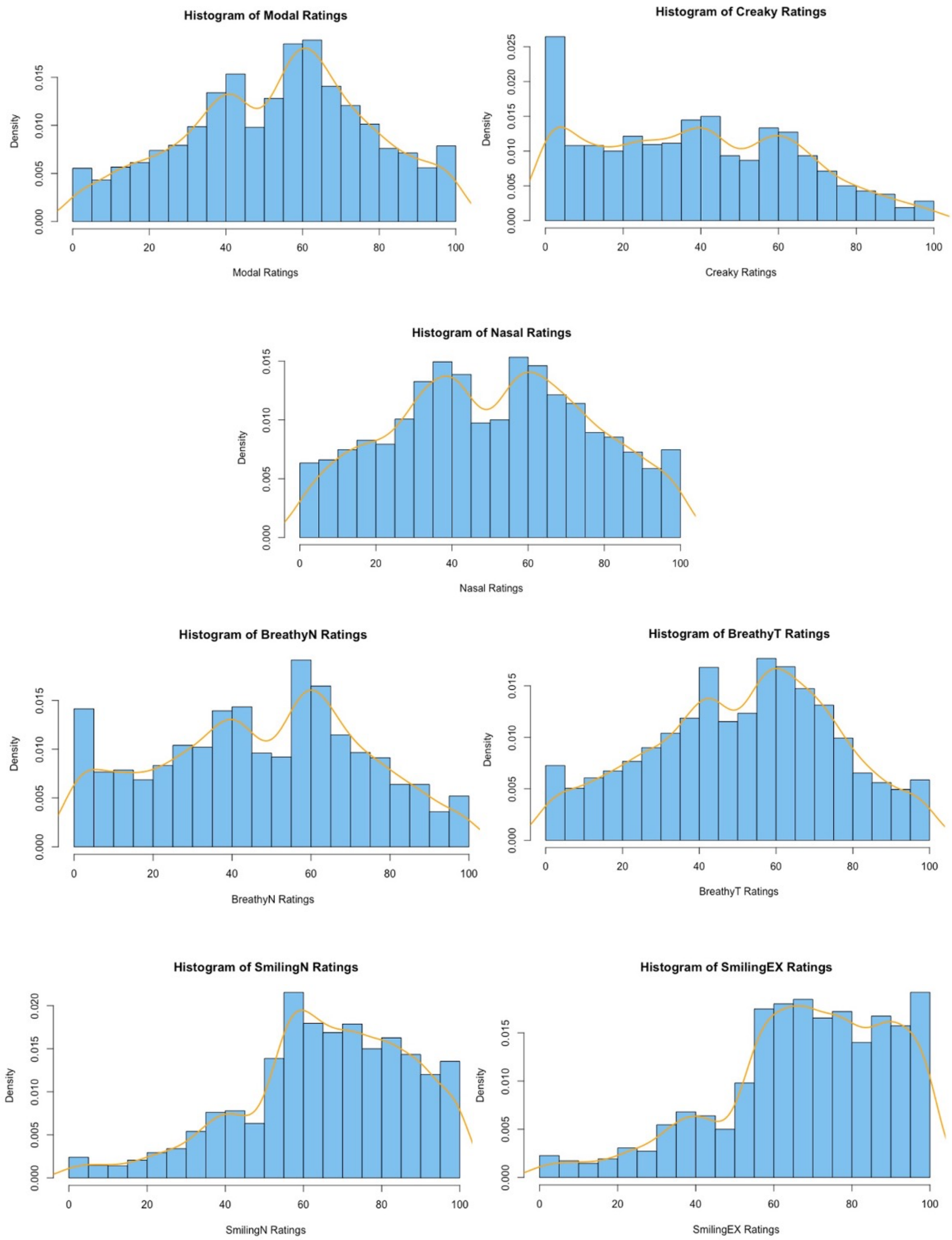


Figure C.1 & C.2: Density plots showing the distributions of the listener responses for the seven examined vocal qualities.

CHAPTER 4

The Influence of Allophonic Variation on the Perceived Charisma of a Speaker

This study has been submitted to the journal *Phonetica* as Pearsell, S., Pape, D., Service, E. (2024). The influence of allophonic variation on the perceived charisma of a speaker.

Abstract

The acoustic properties of a speaker's voice have been shown to impact the perception of their personality. Previous research examined the influences of pitch, pause-filled gaps, gender, etc., on perceived socially important speaker characteristics such as charisma. In contrast, the effect of allophonic variation on perceived personality traits is currently understudied. In this paper, we examine how allophonic variations like (1) the presence versus absence of final consonant devoicing, and (2) substitutions like flapping ([ɾ]) and glottal stop ([ʔ]) for the alveolar [t] production in Canadian English affect the perceived charisma of a speaker. We examined the personality ratings of speakers in two contexts: *in-person* (Experiment #1) and *online* (Experiment #2). Listeners were tasked to rate ten statements selected from the Big 5 of personality traits for each presented audio stimulus. They were rated using continuous sliding scales (*strongly agree* to *strongly disagree)*. For in-person (lab conditions) participants, we found statistically significant rating differences for consonant devoicing: final voiced items were rated higher compared to devoiced ones. For the /t/ variation, we found no statistically significant differences apart from speaker differences.

For online participants, we did not find statistically significant results for final consonant devoicing. For the /t/ variation part, we found statistical significance for [t] productions versus the glottal stop variant, and also for the flap productions versus the glottal stop, with no rating differences between [t] and flap. These results demonstrate that, firstly, segmental differences such as allophonic variant selection affect charisma ratings. Secondly, the choice of the participant platform (in-person vs online) strongly affects the measured effects we find on the perception of allophonic variation and in turn of perceived charisma.

1. Introduction

The acoustic properties of speech which influence the perception of personality traits have received increased research interest in recent years. This research has examined of the acoustic signal in terms of features such pitch (Puts et al., 2007; Rosenberg & Hirschberg, 2009; Quené et al, 2016; Berger et al., 2017), pause-filled gaps (Möbius, 2003; Niebuhr & Fischer, 2019), amplitude or loudness (Novák-Tót et al., 2017), etc., on the personality trait ascription of speakers (e.g., dominance, or charisma). To our knowledge, research on the influence of allophonic variation within English-language speech on personality trait ascription remains limited. Here we investigate the allophonic effects of final consonant devoicing (e.g., /hæd/ becoming /hæt/), as well as plosive variation such as alveolar flapping (e.g., water /wɔɾɚ/) and glottal stop substitution (e.g., water /wɔʔɚ/) on the perceived charisma of a speaker.

1.1 Previous Research

Allophones are variants of a particular phoneme which do not change the meaning of the word and function in systematic ways determined by phonological contexts. An example is /l/ and velarized (dark) /ɫ/ in the English words light (/laɪt/) versus fall (/fɑɫ/). The former is produced by the tongue tip touching the alveolar ridge with approximant constriction. The latter is also produced by the tongue tip touching the alveolar ridge with approximant constriction but with additional raising of the tongue body in the velar region. Allophones are interesting because they occur in complementary distribution, i.e., the phonotactic environment determines which variant is produced. Other factors, such as the geographical region (or dialect) of a particular language (e.g., Toronto, Ontario versus Calgary, Alberta), the sociolinguistic background of a speaker (e.g., upper class versus lower class), and language background (e.g., the spoken language is not a speaker's native language) can impact which allophone is expected and produced. Possible productions include variants which are uncommon or non-standard for a specific variety of spoken language.

The complexity of allophones and their distributions, both across language varieties and within a variety of a language, allows questions about their sociolinguistic function. One such question concerns how allophonic distribution may impact perceptions of charisma. To our knowledge, there are no studies which examine the variation in allophone production in the context of personality trait perception. However, one study by Niebuhr (2017) found systematic effects of charisma ratings as a result of varying degrees of reduction within prosodic and segmental domains. This is systematic effect may be relevant to allophonic variation as this variation also behaves systematically. The study by Niebuhr examined several aspects of speech reduction involving segmental and prosodic components. These included naturally produced reductions within sentences and included a spectrum of (1) fully produced sentences, (2) slightly reduced productions (e.g., informal conversation), and (3) strongly reduced productions (i.e., content and function words were greatly reduced in production). With a specific focus on the segmental results, the study found that the stronger the reduction in the sentence, the greater the probability of listener-perceived speaker attributions of absent-mindedness or clumsiness. The speaker was also perceived as less skilled, less sociable, less optimistic, and less educated. All of these attributes have been linked to a reduced perception of charisma in speakers. In other words, the greater the phonetic reduction in speech production, the less charismatic a speaker was perceived. Although the research by Niebuhr was not directly related to perceptions of allophonic variation, it does demonstrate the systematic effects due to phonemic or prosodic variation and shows these systematic effects may predict the outcomes of speech reductions on perceptions of charisma. These results provide preliminary evidence that variation in speech output, as found in segmental reduction, impacts listeners' ratings of speakers' personality. Similar results could be found for allophonic variations as allophonic variants often have variations which display degrees of reduction particularly from the underlying representation. For example, opting to produce a reduced form of a medial /t/ in North American English (flap [ɾ]) rather than the full articulated [t], or opting to produce an even more reduced form (glottal stop [ʔ]).

1.1.1 Allophonic Variation

Across languages and dialects within languages, the phonological rules which determine the surface realization of a phoneme, or the produced allophone, vary depending on several factors. These include the phonemic inventory of that language and language change over time, as well as sociolinguistic influences such as the speaker's birthplace or age. The two instances of allophonic variation targeted in the current study are final consonant devoicing and voiceless alveolar stop/plosive variation in standard North American English.

Final devoicing is a phonemic neutralization process in which the word-final stop or fricative consonant loses its voicing feature, for instance, [d] becomes [t] in the surface form. Final consonant devoicing does not typically occur in standard varieties of North American English, as demonstrated by the existence of distinct minimal pairs such as *had* [hæd] and *hat* [hæt]. Although this is the general phonological pattern, there are certain dialects of English, such as African American Vernacular English (A.A.V.E.), as well as the English spoken by non-native speakers whose first language contains this phonological process (i.e., German, Polish, Dutch, Russian, etc.). In these languages, the word-final phonologically voiced consonants are phonetically produced as devoiced in their surface forms. Although research has documented and examined the phonological process of final consonant devoicing in various English dialects, little is known about the impact of this variation in Native English speakers (specifically Native Canadian English Speakers) on personality trait attribution.

The second variation studied here involves voiceless alveolar stops/plosives (/t/). In English, particularly Canadian English, phonological variants of the underlying phoneme /t/ can occur in several environments. Environments for the phoneme /t/ manipulated in the current study were i) between two vowels (including syllabic sonorants; Walker, 2015), ii) following a nasal and before an unstressed vowel, iii) across word and syllabic boundaries when /t/ is in the coda position of the first word/syllable and an unstressed vowel is in the onset position of the second word/syllable. Essentially, /t/ cannot be in a stressed position in order for a different variant to be produced. In the current study, the specific /t/ allophones for these environments were the alveolar flap ([ɾ]; also called *flapping*) and the glottal stop/plosive ([ʔ]; also called *t-glottalization*). Although other allophonic variants have been proposed for these environments, such as deletion, [ɾ] and [ʔ] are both more consistently measured and agreed upon (Walker, 2015; Patterson & Connine, 2001). To clarify, Table 1 demonstrates these environments, as well as an example and transcriptions of possible variants of interest for this study.

Table 1: The environments of interest, an example word and an IPA transcription of the three possible outcomes of production. These transcriptions demonstrate the three allophone variants of interest.

Environment	Examples(s)	IPA Transcription
(i) between vowels: a) when second vowel is unstressed (within word)	city	[sɪ**t**i] [sɪ**ɾ**i] [sɪ**ʔ**i]
b) including syllabic consonants	little	[lɪ**t**l̩] [lɪ**ɾ**l̩] [lɪ**ʔ**l̩]
(ii) following a nasal and preceding an unstressed vowel -	wanted	[wɑn**t**əd] [wɑn**ɾ**əd] [wɑn**ʔ**əd]
(iii) between vowels (second vowel is unstressed) – across word/syllable boundary when /t/ is in coda position of the first syllable	my cat is grey	[maɪ kæ**t** ɪz gɹeɪ] [maɪ kæ**ɾ** ɪz gɹeɪ] [maɪ kæʔ ɪz gɹeɪ]

Allophonic variation takes many shapes because of varying phonological rules both in and across languages. Examination of how allophonic variation influences personality trait perception can provide interesting insights into how speakers' linguistic manifestations can impact perceptions of charisma. This is relevant for both native speakers, as well as second language learners of English who may have different phonemic inventories and phonological patterns resulting in a non-standard allophone production. Additionally, it can provide insight into the productions in a variety of environments such as informal or rapid speech.

1.1.2 Personality Traits and Charisma

One definition of charisma offered by Niebuhr et al. (2019) states: "*charisma is symbolic, emotional laden, and value-based communication style signaling leadership qualities such as commitment, confidence, and competence that affect followers' beliefs and behaviours in terms of motivation, inspiration, and trust.*" Previous research has attempted to conceptualize charisma and charismatic speech by examining perceptual ratings by listeners for speakers' voices. Often, these ratings were collected through a series of statements or questions correlating to traits associated with charisma. Generally, these statements or questions were mapped to the categories found within the Big 5 of Personality Traits (Norman, 1963; McCrea & John, 1992), a model of the structure of personality based on psychometric test instruments. Scores of these instruments allow the identification of five different personality traits: openness, conscientiousness, extroversion, agreeableness, and neuroticism (O.C.E.A.N.).

Research by Rosenberg and Hirschberg (2009) demonstrated that *charismatic speakers* were associated with the personality traits of being enthusiastic, charming, persuasive, and convincing, all traits that fall into the categories within the Big 5 of personality model. Antonakis et al. (2016) found that confidence and self-assuredness are linked to extroversion and charismatic speech. However, Pearsell & Pape (2023) pointed out the that "label" of a charisma relies on vague interpretations or definitions and thus fail to sufficiently explain exactly what charisma in a speaker means. Alternatively, relying on *the Big 5* of personality traits as a construct for explanation provides a mapping which is too broad to capture charisma, as all facets of personality traits can be sorted into one of the five categories within *the Big 5*. In search of more consistent scores, Pearsell and Pape went on to operationalize the concept of *charisma* by tying the ratings of human voices to a collection of statements reflecting the Big 5 factors. They examined the connection between charisma perception based on speech features (specifically voice quality) and personality trait ratings within the paradigm of the Big 5, following the methodology and results of previous research (Rosenberg & Hirschberg, 2009; Tshkay et al, 2018). Following the research by Rosenberg and Hirshberg (2009) and Tshkay et al. (2018) these traits were extroversion, agreeableness, conscientiousness, and neuroticism. Openness was omitted as it was found to not correlated to ratings of charisma. By analyzing *charisma* within the four traits of the Big 5 found to correlate to charisma, Pearsell & Pape developed a clearer and more concrete interpretation of *charisma* as a concept. Rather than identifying single Big 5 factors related to charisma, their study demonstrated that ratings of personality traits *all* increased or decreased together, depending on the positive or negative perception of a specific vocal quality. Increases/decreases of each investigated personality trait perceptions were linked with charisma. This means that the higher, more positive, the ratings of traits, the greater the perception and saliency of charisma, whereas the lower, more negative the ratings, the lower the perception and saliency of charisma (for neuroticism, the scale was inverted). The pattern across personality traits with all traits either increasing or decreasing together, this suggests that not all personality traits (of the *Big 5*) need to be utilized in experimental designs to capture meaningful interpretation in terms of the presence of charisma.

Following this methodology, the current study aimed to understand how a set of ratings of personality traits considered together reflected the perceived charisma of speakers in relation to allophonic variation. To accomplish this, speakers' productions of the different allophonic variants of interest were rated by listeners on scales constructed to detect the *Big 5 of Personality Traits*. The combination of positive ratings from these scales (i.e., the personality traits of the *Big 5*) was used as an operationalization of the definition of *charisma*.

1.2 Aims and hypotheses

The present study investigates how the choice of certain allophones (i.e. final consonant devoicing and /t/ variation) influences listener ratings of speakers' personality traits. It also investigates the correlation between results collected from separate groups of participants on two different platforms, *online* versus *in-person* (i.e., *in-laboratory*), respectively. The study was planned to consist of two parts: a final consonant devoicing block and a /t/ variation block. For the final consonant devoicing component, three variants were used: (i) *voiced* (final consonants were voiced), (ii) *devoiced* (final consonants were devoiced), and (iii) *neutral* (no environments for final consonant devoicing were present in the stimuli). For the /t/ component, three allophonic variants were selected: (i) [t] (voiceless alveolar stop/plosive), (ii) [ɾ] (alveolar flap/tap), (iii) [ʔ] (glottal stop/plosive).

We formulated the following hypotheses:

1. Because maintaining voicing on final consonants is the typical allophone production in Canadian English, we expected that for the particular set of stimuli selected for this component, maintaining the voicing on word-final consonants would result in the most positive rating in relation to personality traits (i.e. together reflective of charisma). Therefore, we expected the set of voiced final consonants present in this study (i.e., final /v d z g/) to result in higher (i.e. more positive) ratings for all personality traits, and the set of devoiced conditions (i.e., final /f t s k/) to produce lower (i.e. more negative) ratings. A neutral condition was also present in this component. Neutral conditions would generate the same ratings as voice conditions and, in turn, devoicing conditions would rate lower than neutral ones.
2. Because the full alveolar [t] surface form is the standard variant for Canadian English population for the /t/ variation component, we, expected it to result in the highest (most positive) personality trait ratings, and thus would define the baseline of a speaker's personality judgment. The [ʔ] condition was expected to produce the lowest (most negative) ratings, as it is the least frequent variant of this phoneme for our population (Walker, 2015; Patterson & Connine, 2001). Patterson and Connine (2001) demonstrated that [ɾ] is the most frequently produced allophone of /t/. This could result in one of two outcomes: either the ratings of the [ɾ] condition to fall between the ratings of [t] and [ʔ], as it is the most frequent allophone of /t/, despite not being the theoretical standard production (Patterson & Connine, 2001). An alternative potential outcome could be that [ɾ] would result in the highest personality ratings due to its frequency. It would then rate higher, or more positively, than [t] which is the underlying representation.

2. Experiment 1: In-Person

This experiment aimed to find the effects of allophonic variation for final consonant devoicing and /t/ variation in flapping environments on the perception of charisma. This experiment focused on listeners participating in-person.

2.1 Methods

2.1.1 Participants

This was an *in-person* experiment with 20 participants ranging in age from 18 to 24 (mean = 19.45; s.d. = 1.50) all of whom were undergraduate students at McMaster University, recruited using the SONA experiment participation platform. All participants reported normal cognition and hearing, and all answered a set of demographic questions regarding age, gender, acquired and spoken languages, as well as musical education background.

Participants were presented with all allophonic variants by each speaker. As a result, the total duration of the experiment, including pre-screening components, was approximately 60 minutes and participants received course credit. Ethics clearance was obtained from the McMaster Research Ethics Board (MREB).

2.1.2 Stimuli

We created five paragraphs for the final consonant devoicing component and three paragraphs for the /t/ variation component. Each paragraph consisted of 5 simple sentences as the basis of the language material to be used as stimuli in the perception study. These paragraphs were constructed using words with neutral emotional valence (i.e., the emotional connotation invoked by these paragraphs was neither positive nor negative) and avoiding words perceived to have high emotional valence to prevent any influences of positive or negative valence on listeners' interpretations of the speaker's voice. Each paragraph was approximately 8 seconds long.

The experiment was separated into 2 components, a devoicing variation component, and a /t/ variation component. The recorded stimuli for both components were produced by six native Canadian English speakers (3 male, 3 female). Speakers recorded the stimuli using a high-quality microphone (Sennheiser ME63) and Focusrite Scarlet audio interface in the certified sound-proof booth of the ARiEAL Phonetics Lab at McMaster University. For all recordings, the microphone distance was specified to be around 10 cm, with the microphone being horizontally off-center (from the lips) by 30–45 degrees.

The speakers recorded each paragraph three times for each version of both final consonant devoicing and /t/ variations. The best of these productions was selected as the perceptual stimuli for the experiment. The criterion for the best production was based on the least amount of production errors, the most neutral f0, the most neutral amplitude, and the most natural speech rate and other prosodic patterns. Differences in prosody were controlled as tightly as possible across conditions and speakers. After recording, each audio sample was screened and checked carefully by the two authors of this study using the audio editor Amadeus Pro (Hairer, 2021) for (speaker-produced) accuracy and consistency. This included the production of vocal intensities, differences in articulation (e.g., different speech rates and hypo/hyperarticulation), as well as the lack of undesired voice quality changes (e.g., sentence-final glottalization). After monitoring and selecting the best production, we normalized loudness for all stimuli. For this normalization, all stimuli (i.e., all paragraphs for all speakers) were normalized for loudness to 70 dB mean intensity using Praat software's (Boersma & Weenink, 2023) intensity normalization procedure.

2.1.2.1 Final Consonant Devoicing

The current study explored the effects of two kinds of allophonic variation: final consonant devoicing and /t/ variation. For the *devoicing component* of the experiment, five paragraphs were produced by human speakers. Three of these five paragraphs contained three different instances of potential final devoicing environments. Two paragraphs contained no instances of potential final devoicing. These were used as distractors/fillers. Thus, 3 different conditions were defined here: the first (voiced) and second (devoiced) conditions used the potential final devoicing paragraphs in which speakers produced *all* the devoicing instances as voiced for the first condition, and *all* as devoiced for the second condition. The third (neutral) condition for this component consisted of 2 paragraphs with no instances of possible final devoicing (thus no final stops or fricatives). This resulted in 48 acoustic stimuli (6 speakers x 3 paragraphs x 2 final devoicing conditions + 6 speakers x 2 distractor/filler paragraphs).

2.1.2.2 /t/ Variation

For the /t/ *variation* part of the experiment, three paragraphs were produced by speakers, each containing three instances of flapping environments. Speakers produced each of these paragraphs in three conditions: with clear and fully articulated voiceless alveolar stops ([t]) for the first (voiceless) condition, with glottal stops [ʔ] (as substitution of the alveolar stop) for the second (glottal stop) condition, and as alveolar flaps [ɾ] (as substitution of the alveolar stop for the third (flap) condition). This resulted in 54 acoustic stimuli (6 speakers x 3 paragraphs x 3 /t/ variant conditions) for this component of the experiment. Overall, 102 paragraph stimuli were used for the entire experiment.

2.1.3 Experimental Setup

The experiment was conducted in a certified soundproof room within the Phonetics Lab at McMaster University and used state-of-the-art acoustic playback equipment (Focusrite Scarlett audio interface, Sennheiser HD 598 linear frequency-response headphones). The authors used the Gorilla Experiment Builder to create and host the experiment.

With each speech sample, participants were presented visually with 10 statements on a computer screen alongside a randomized acoustic stimulus. The statements used were from a previous study by Pearsell and Pape (2023) and had been initially modelled after the two studies by Rosenberg and Hirschberg (2009) and Tskhay et al. (2018). As Pearsell and Pape explain in their earlier study, the selected statements were designed to reflect four of the five traits from the *Big 5 of Personality*. The fifth trait, *openness*, had not been found to be related to charisma ratings and was therefore omitted (Rosenberg &

Hirschberg, 2009; Tskhay, 2018). We inverted the scores for neuroticism of the traditional *Big 5*[13] to be in line with expected positive correlations with *charisma*.

As the intention of the researchers was to establish a relationship between the speaker and the listener, the speakers were depicted as professors[14] (see Figure 1 below). By framing this relationship as one between listener and professor, the authors wanted to prevent any other possible social interpretations with the participants (who were undergraduates).

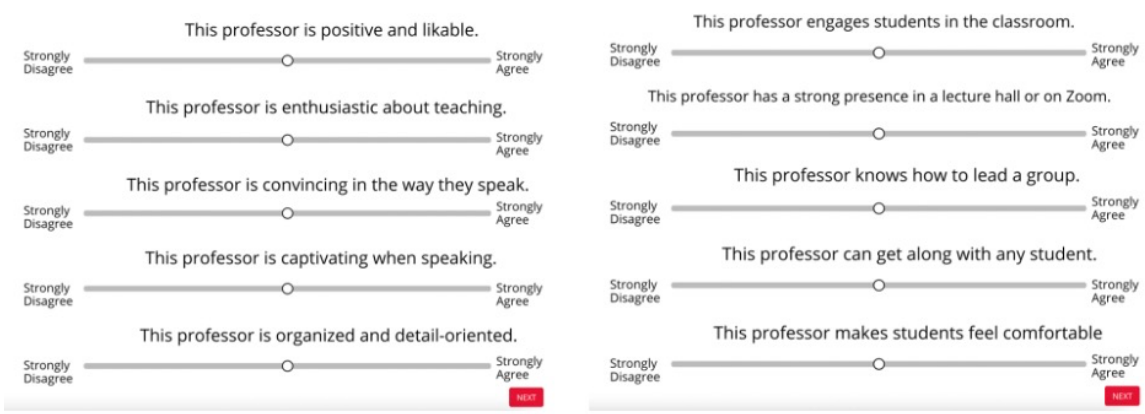


Figure 1: Screenshot of the sliding scales with the questions presented to the listeners during the perception study: shown is the screen with the first five questions (left panel) and the second screen with the last 5 questions (right panel).

The 10 statements were split into two screens, with the first screen of the experiment containing the first 5 statements and the following screen presenting the last 5 statements to prevent any visual or cognitive overload for participants. Each auditory stimulus was played once per screen (i.e., repeated once for the second screen), and only occurred once during the experiment. Participants would slide the "button" on a continuous scale to the desired location for each statement to show the degree to which they either agreed or disagreed with each presented statement. The responses on these scales gave a rating from 0% (*strongly disagree*) to 100% (*strongly agree*).

2.1.4 Statistical analysis

Statistical analysis was performed using the software R (R Core Team, 2018) and RStudio (R Studio Team, 2018). The listener ratings of personality variables on continuous scales were examined as the dependent variable. The statements for each

[13] The *neuroticism* personality trait score was inverted as lower original scores for neuroticism are higher in participants' ratings for charisma. In the traditional rating, higher scores for this trait positively correlate with more negative connotations, such as proneness to negative emotions, or anxiety; the higher the *neuroticism* score, the more the trait is displayed; the lower the score, the less the trait is displayed.
[14] This relationship was established as all participants for the in-person sample were students.

included trait corresponding to *the Big* 5 were first averaged for each individual trait after which these four resulting averages for each trait were then averaged, again, into a final, singular average. A repeated measures ANOVA was used to analyze these final mean scores. Tukey's post hoc test was used for comparisons between speakers and allophonic variants for each component of the experiment.

2.2 Results

2.2.1 Final Consonant Devoicing

Figure 2 shows great variability in the mean ratings of the examined allophonic variants for final consonant devoicing by in-person participants. For five of the six speakers (except for the speaker coded as BF), the devoiced final consonant variants were rated the lowest (the most negative) for the perception of charisma. For three speakers (CA, CG, MK) the voiced condition was rated the most positive. For the remaining three speakers (BF, GR, MC) the neutral condition was rated the highest. It can also be seen that the difference between voiced and devoiced is a little more pronounced for most speakers than the difference between voiced and neutral, where in principle no differences would be expected.

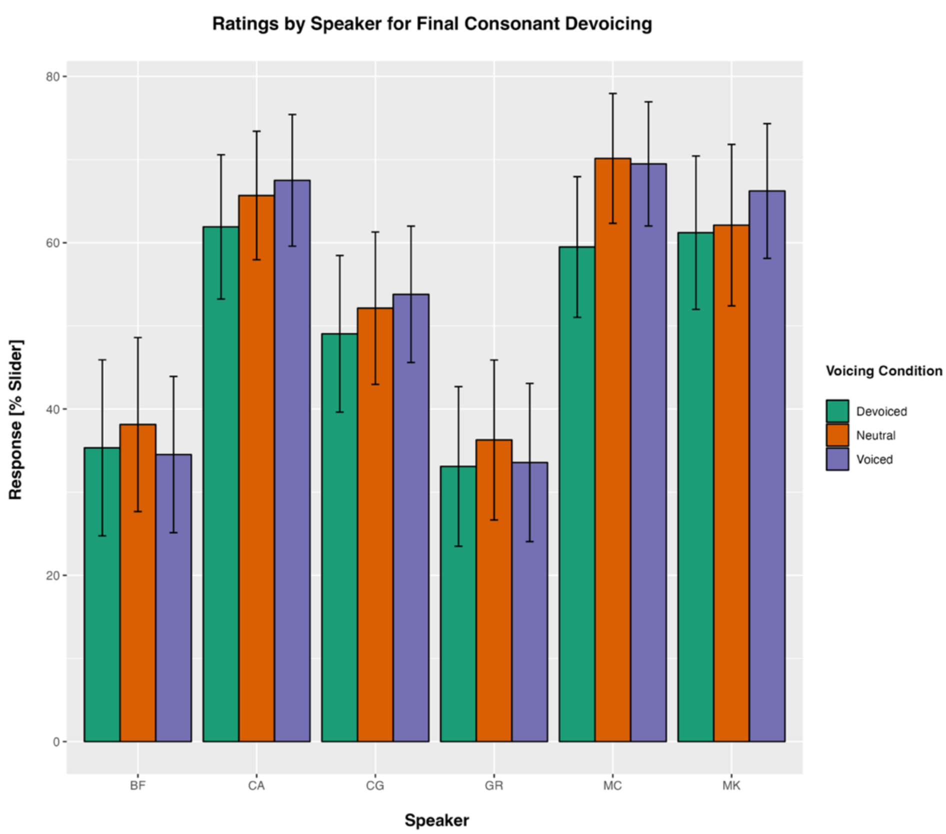


Figure 2: Mean charisma ratings for different allophonic variants of final consonants by all in-person listeners shown separately for the six speakers' productions. The y-axis displays the slider position

percentage with 0% corresponding to the left extreme value and 100% as the right extreme value of the slider. The x-axis shows the different speakers, and the coloured bars the different allophonic variants (devoiced, neutral, voiced). The error bars are standard errors.

A repeated measures 6 x 3 ANOVA was performed to examine whether charisma ratings were reliably affected by final consonant devoicing. The independent variables were speaker (six speakers) and voicing (voiced, devoiced or neutral). The results from the repeated measures ANOVA found that speaker identity ($F(5, 95) = 68.13$, $p < 0.001$, $\eta^2_p = .78$), voicing condition ($F(2, 38) = 11.40$, $p < 0.001$, $\eta^2_p = .38$), and the interaction between speaker and voicing ($F(10, 190) = 3.16$, $p < 0.001$, $\eta^2_p = .14$) all significantly affected ratings of personality. Unsurprisingly, individual speakers were rated as differing in charisma. Tukey's post hoc difference (Table 2) tests also indicated statistical significance between devoiced and voiced conditions, as well as devoiced and neutral conditions, with lower ratings for the devoiced variant than the two other variants of final consonant production. No statistical significance was found between voiced and neutral conditions.

Table 2: Post hoc test of the main effect of final consonant devoicing on charisma ratings for in-person participants.

Comparison							
Voicing		**Voicing**	**Mean Difference**	**df**	**t**	**p**	**p_{tukey}**
Voiced	-	Devoiced	4.141	19.0	3.512	0.002	0.006**
Voiced	-	Neutral	0.049	19.0	0.071	0.944	0.997
Devoiced	-	Neutral	-4.092	19.0	-3.899	< .001	0.003**

As the repeated measures ANOVA indicated statistical significance for the interaction between speaker and voicing, Tukey- corrected post hoc analysis was carried out. It revealed that the speaker MC was judged as less charismatic in the devoiced compared to the voiced and neutral conditions, whereas the five other speakers where not rated as statistically significant in their differences amongst the conditions (Table 3). Due to the high number of tests, there was not enough power to detect other significant differences between conditions. However, it can be seen in the uncorrected column of *p*-values that the ratings for the other speakers (CA and MK) showed the same tendency, contributing to the main negative effect of devoicing. (Table 3, Figure B1).

Table 3. The results of the post hoc analysis for the interaction between devoicing and speaker for in-person participants. Results are separated by speaker and allophonic variant. Columns show the results for mean difference, t score, the uncorrected p-value and the p-value adjusted with Tukey correction.

Speaker		**Mean Difference**	***t***	***p-value* uncorrected**	***p-value* Tukey adjusted**
BF	Voiced vs devoiced	-0.91	-0.64	0.529	1.000
	Voiced vs neutral	-3.70	-2.04	0.055	0.817
	Devoiced vs neutral	-2.80	-1.53	0.143	0.976
CA	Voiced vs devoiced	5.61	2.61	0.017*	0.491
	Voiced vs neutral	1.83	0.92	0.367	1.000
	Devoiced vs neutral	-3.78	-1.55	0.138	0.973
CG	Voiced vs devoiced	4.75	2.07	0.053	0.805
	Voiced vs neutral	1.43	0.72	0.483	1.000
	Devoiced vs neutral	-3.32	-1.39	0.181	0.990
GR	Voiced vs devoiced	0.47	0.30	0.770	1.000
	Voiced vs neutral	-2.71	-1.49	0.151	0.980
	Devoiced vs neutral	-3.18	-1.56	0.136	0.972
MC	Voiced vs devoiced	9.90	4.67	< 0.001***	0.013*
	Voiced vs neutral	-0.66	-0.29	0.776	1.000
	Devoiced vs neutral	-10.56	-4.65	< 0.001***	0.013*
MK	Voiced vs devoiced	5.02	2.79	0.012*	0.389
	Voiced vs neutral	4.11	2.73	0.013*	0.420
	Devoiced vs neutral	-0.91	-0.44	0.666	1.000

2.2.2 /t/ Variation

The mean rating scores for the /t/ variation component of the experiment are shown in Figure 3. Visually, there appears to be a lack of general trends for differences between allophonic variants. For some speakers (CG, MK), the glottal stops received the most positive ratings for charisma while for others (MC) these variants received the most negative ratings or ratings at par with the voiceless alveolar plosive (BF, GR). Based only on a visual investigation, there does not appear to be a consistent relationship between the allophonic variants and ratings of charisma. Ratings appear to be more speaker-dependent than variant-dependent.

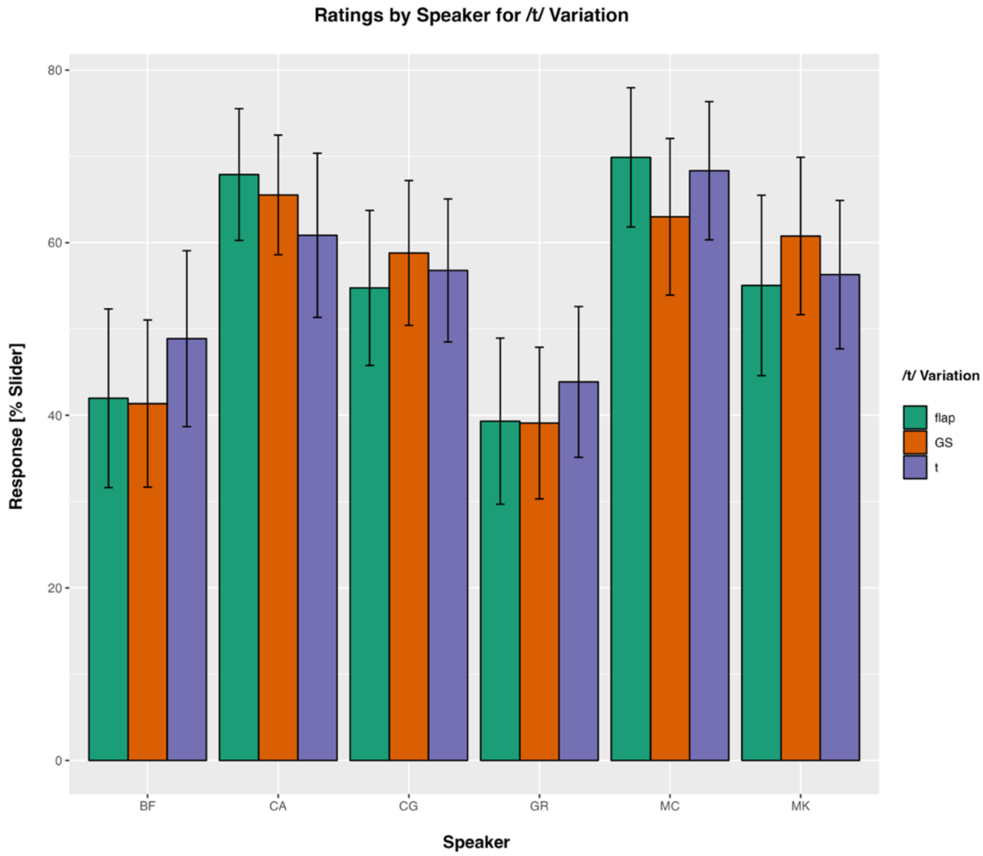


Figure 3: Charisma ratings for different /t/ variants by all in-person listeners for the six speakers' productions. The y-axis displays the slider position percentage with 0% corresponding to the left extreme value and 100% as the right extreme value of the slider. The x-axis shows the different speakers, and the coloured bars the different allophonic variants: flap([ɾ]), glottal stop (GS; [ʔ]), and t ([t]). The error bars are standard errors of the mean.

A repeated measures 6 x 3 ANOVA was performed to examine how each /t/ variant affected listeners' ratings for personality traits associated with charisma. For this component, the independent variables were speaker and /t/ variant (i.e., flap [ɾ], glottal stop [ʔ]), and voiceless alveolar plosive [t]). The results from the repeated measures ANOVA revealed that speaker ($F(5, 95) = 48.14$, $p = < 0.001$, $\eta^2_p = 0.717$) and the interaction of speaker and /t/ variant ($F(10, 190) = 7.20$, $p = < 0.001$, $\eta^2_p = 0.275$) significantly affected personality ratings. No statistical significance was found for the main effect of /t/ variant ($F(2, 38) = 0.72$, $p = 0.494$, $\eta^2_p = 0.036$).

Tukey's post hoc tests were used to explore the interaction between speaker and /t/ variant. Pairwise comparisons (Table 4) show that the glottal stop variant was associated with significantly lower charisma ratings than /t/ for speaker BF. No other pairwise comparisons were statistically significant, suggesting that the interaction between speaker and /t/ variant was driven by trends for opposite preferences for different speakers (Figure 3). Like the results for the devoicing component, there was not enough power to reliably detect preferences for each speaker due to the high number of tests performed. This is an interesting result as allophones clearly mattered but affected ratings differently for different speakers.

Table 4: The results of post hoc tests based on the interaction between speaker and /t/ variant on personality ratings by in-person participants. Results are separated by speaker and allophonic variant (flap [ɾ], glottal stop (GS; [ʔ]), and voiceless alveolar plosive (standard) [t]. Columns show the results for mean difference, *t* value, the uncorrected *p-value* and the *p-value* adjusted with Tukey correction.

Speaker		Mean Difference	*t*	*p-value* uncorrected	*p-value* Tukey adjusted
BF	[t] vs flap	6.91	4.38	< 0.001***	0.023*
	[t] vs GS	7.27	2.37	0.029*	0.635
	flap vs GS	0.36	0.17	0.870	1.000
CA	[t] vs flap	-7.04	-3.08	0.006**	0.254
	[t] vs GS	-4.67	-2.32	0.032*	0.666
	flap vs GS	2.38	1.69	0.107	0.947
CG	[t] vs flap	1.85	1.21	0.239	0.997
	[t] vs GS	-1.93	-1.30	0.208	0.995
	flap vs GS	-3.78	-2.11	0.048*	0.782
GR	[t] vs flap	4.55	2.79	0.012*	0.388
	[t] vs GS	4.77	2.41	0.026*	0.610
	flap vs GS	0.22	0.13	0.895	1.000
MC	[t] vs flap	-1.51	-0.84	0.414	1.000
	[t] vs GS	5.38	2.91	0.009**	0.327
	flap vs GS	6.89	3.58	0.002**	0.113
MK	[t] vs flap	1.25	0.66	0.514	1.000
	[t] vs GS	-4.47	-2.39	0.028*	0.623
	flap vs GS	-5.72	-2.40	0.026*	0.611

2.3 Discussion and Conclusions

The current experiment tested the hypothesis that the selection of phonetic variants in a person’s speech could affect how charismatic they are perceived to be, all for in-person listeners participating in a laboratory setting. We tested two cases of allophonic variation. Our results demonstrate systematic effects for final consonant devoicing in English. Final consonant voicing versus devoicing leads to statistically significant and consistent effects on personality trait evaluations. However, the /t/ variation did not show the same statistically significant results.

We found that maintaining the voicing of final consonants leads to higher, and thus more positive, personality trait ratings on scales contributing to the perception of charisma compared to devoicing the consonants, thus confirming our first hypothesis that Canadian English listeners prefer speakers who maintain final consonant voicing, matching the allophonic distribution of final consonant voicing in standard North American dialects. Opting to produce the underlying variant (i.e., maintaining voicing in final consonant positions) increased positive personality trait perception. Devoicing the final consonant decreases ratings of personality traits reflecting charisma, at least in standard North American varieties of English. As expected, we did not find differences between the voiced consonant condition and the neutral (i.e. no final voicing items present) condition. The present findings for final consonant devoicing appear to fall in line with the outcomes found in Niebuhr (2017), who examined the effects of reduction in prosodic and segmental domains. In that study, the outcomes displayed systematic effects of reduction and thus correspond to the systematic effects found in our allophonic distributions for devoicing patterns.

In the /t/ variation experiment, we found significant effects for one speaker but inconsistent trends amongst the other five. Although the main effect was not significant, effects within speakers suggest that ratings for speech with different /t/ variants are more speaker dependent than general for specific allophone selection. This could be a result of other factors within a speaker's voice, such as timbre or other qualities, "colouring" the effects of /t/ variants. The lack of significant main effects goes against our second hypothesis in which we anticipated the standard variant of Canadian English, [t], to produce the highest (most positive) personality trait ratings whilst [ʔ] condition was expected to produce the lowest (most negative) ratings, as it is the least frequent variant of this phoneme for our population. Thus, we reject our second hypothesis: the choice of variant of /t/ does not appear to have consistent effects on personality trait ratings reflecting the perception of charisma, at least not generally at the level of allophone. However, it may play a role on speaker level.

As a more general takeaway, our results indicate that some allophonic variations, such as final consonant devoicing, play a role in personality trait attribution, whereas others, such as /t/ variation, have no consistent effects in North American English in our data. It is important to note that although the speaker effect was very strong (as can be seen in the main plots and the interaction plots in Appendix C, Figure 1), there was still a significant effect of some phonetic variables. Opting to produce the non-standard devoicing in the final consonant position led to decreases in the perception of charisma whereas the standard variant utilized in the /t/ variation appeared to affect the charisma perceptions of different speakers in different ways. The reason for this is speculative at this point, requiring further research and possibly a larger range of allophonic variants for examination. Regardless, this research provides preliminary findings for understanding how allophone selection and usage impact perception of personality traits and charisma.

As there was an interaction between phonetic variant and speaker, as well as increased variation among raters, there is a possibility that the raters' own use of allophonic variants may have played a role while affecting their ability to identify with the speaker. Such variables could be controlled in future studies.

An exciting perspective is that this research could be extended to the field of Artificial Intelligence (AI). With the increased usage of AI in our daily lives, understanding how to utilize the speech capabilities of AI (i.e. synthesis procedures) is an important task. Previous research has demonstrated (Fischer et al., 2019) that, much like humans, AI also benefits from using charismatic speech techniques, such as variation of f0 range, or tempo, which have been found to elicit perceptions of a speaker being more persuasive and influential.

We recognize several limitations of this study. Firstly, there was only a small number of recorded speakers for the stimuli generation. Although more speakers would have been preferable, having more speakers would have resulted in longer experimental run time. With longer experiments comes the increase of cognitive workload which could have had detrimental effects on the reliability of the results. Although the number of rated speakers could be increased, we do believe that the current number of speakers provides interesting insights into the impact of allophonic variation.

As previously mentioned, another limitation of the study is the cross-speaker variation which is problematic when using naturally produced stimuli (see Appendix A). However, to have the most natural sounding stimuli without potential artifacts from software manipulation, naturally produced stimuli with as tightly controlled productions as possible are preferable.

Finally, the established relationship between the listeners and the speakers could also play a role. In the current experiment speakers were labelled as professors. This established a listener-speaker relationship which may have affected how allophonic variations were perceived. Perhaps, the results would have been different if the relationship between listener and speakers had been defined differently (e.g., the speaker labelled as a student, or friend, or labelled as a co-worker, etc.). There is also a chance that results could be different if the description of the social environment had been different (e.g., socializing with friends, or at work). This could also be true for different listener platforms or modalities, for example, in-person speech interactions versus online communication.

3. Experiment 2: Online

This experiment focused on listeners participating online. It explored whether the effects of allophonic variation for final consonant devoicing and /t/ variation in flapping environments on the perception of charisma could be replicated on a different platform.

3.1 Methods

3.1.1 Participants

74 participants took part in the experiment, recruited from Prolific, an online experiment recruitment platform. Participants were recruited from a pool of self-reported North American native English speakers. They ranged in age from 18 to 71 (mean = 33.36; s.d. = 11.86). All participants reported normal cognition and hearing, and all answered a set of demographic questions regarding age, gender, acquired and spoken languages, as well as musical education background. The experiment was conducted on the participant's personal computer using their own headphones/earbuds.

3.1.2 Stimuli

The same stimulus paragraphs were used as in Experiment 1 (5 paragraphs for final consonant devoicing and 3 paragraphs for /t/ variation). The experiment was separated into 2 components with an identical structure to Experiment 1

3.1.3 Experimental Setup

The experimental setup for this experiment was the same as Experiment 1. Identical to Experiment 1, alongside the auditory stimuli participants were presented visually with 10 statements for rating. Participants rated each of the stimuli on the scales related to the 10 statements with the responses on the scales providing a rating from 0% (*strongly disagree*) to 100% (*strongly agree*).

The total duration of the experiment, including pre-screening components, was approximately 20 minutes. Participants received financial compensation. Ethics clearance was obtained from the McMaster Research Ethics Board (MREB).

Similar to Experiment 1, the experiment was created and hosted on Gorilla Experiment Builder. Data were collected from December 17, 2022, to January 16, 2023.

Unlike in Experiment 1, participants were exposed to only one allophonic variant by each speaker to prevent potential bias from them hearing (and thus evaluating) the same speaker in more than one allophonic condition[15]. Each participant only heard one allophone variant per speaker for each type of variation, based on a Latin square design. For example, listener 1 would hear only the devoiced stimulus paragraphs for speaker 1, voiced stimulus paragraphs for speaker 2, neutral stimulus paragraphs for speaker 3, etc., while listener 2 would only hear the voiced stimuli for speaker 1, the neutral stimuli for speaker 2, and the devoiced stimuli for speaker 3, etc. Online recruitment allowed the opportunity to increase the number of participants due to the ability to run participants both simultaneously and independently of any time constraints of the researcher.

[15] In-person (Experiment 1) participants heard all variants from all speakers. This was due to the limited number of participants available from the SONA participant pool preventing the large pool of participants needed to perform a Latin square design as is present in the online design (Experiment 2).

3.1.4 Statistical analysis

Statistical analysis for this experiment was similar in structure to Experiment 1. The analysis was performed using the software R (R Core Team, 2018) and RStudio (R Studio Team, 2018). The listener ratings of personality variables on continuous scales were examined as the dependent variable. Because all participants had not been exposed to all conditions to avoid the same speaker being heard producing different allophone variants of the same stimulus, observational units were created from triplets of listeners. Together the participants in each triplet were presented with all the condition combinations. In the analysis, the triplets were treated similarly to participants in repeated analyses. The statements were averaged like in Experiment 1, by computing a mean score for each individual personality trait for each triplet, then averaging again over the four traits for a mean final score. A 6 (speaker) x 3 (allophonic variant) repeated measures ANOVA was used to analyze the data (i.e., the mean final scores). Tukey's post hoc tests were used for pairwise comparisons in each component of the experiment.

3.2 Results

3.2.1 Final Consonant Devoicing

Figure 4 shows the mean personality ratings by online participants related to the examined allophonic final consonant devoicing variants. For half of the speakers (CG, MC, MK), the devoicing condition appears to have been rated the lowest (most negative) for charisma while for the other half of the speakers, the voiced condition the appears to have been rated the lowest (BF, CA) or at par with the devoiced condition (GR). Half of the speakers have the highest (most positive) ratings for charisma in the neutral condition (CA, GR, MC) and the other half of the speakers in either the devoiced (BF) or the voiced condition (CG, MK) rated highest. Unlike for in-person participants in Experiment 1, for online participants, final consonant devoicing does not appear to have produced consistent results for each variant.

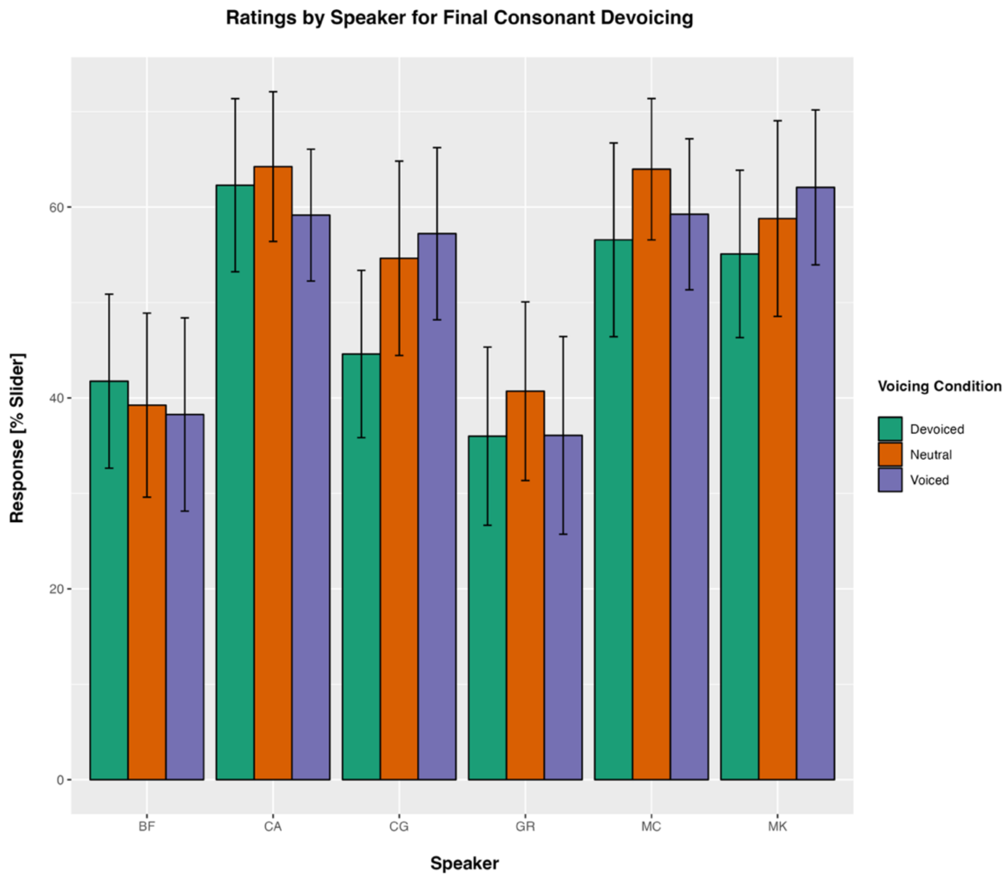


Figure 4: Mean personality ratings by online participants for three final consonant devoicing variants in the 6 speakers' productions. The y-axis displays the slider position percentage, with 0% corresponding to the left extreme value and 100% as the right extreme value of the slider during participant ratings. The x-axis shows the stimuli from different speakers, and the coloured bars the different allophonic variation levels (devoiced, neutral, voiced). Error bars are standard errors.

Like in Experiment 1, a repeated measures 6 x 3 ANOVA was performed to examine how each final consonant devoicing variant affected listeners' ratings of charisma. For this experiment, the independent factors were speaker (stimuli from six speakers) and voicing (i.e., voiced, devoiced or neutral). Unlike in Experiment 1, in the current experiment, listeners were presented with only one variant of the voicing condition from each speaker and the analysis was carried out on triplets of participants who together has experiences all independent variable combinations. There were 21 triplets, resulting in similar statistical power as for the 20 participants in Experiment 1. The results from the repeated measures ANOVA found statistical significance for speaker ($F(5, 100) = 85.60$, $p = < 0.001$), and voicing ($F(2, 40) = 3.84$, $p = 0.030$). No statistical significance was found for the interaction between speaker and voicing ($F(10, 200) = 0.855$, $p = 0.576$).

Tukey's post hoc difference tests for the factor of voicing (Table 5) indicated significantly higher personality ratings in the neutral compared to the devoiced condition. However, no statistical significance was found between voiced and neutral conditions, and, more importantly, there was no significance between voiced and devoiced condition. A non-significant trend for higher ratings for voiced than devoiced was also detected.

Table 5: Post hoc tests of personality ratings in final consonant devoicing conditions for online participants.

Comparison							
Devoicing		**Devoicing**	**Mean Difference**	**df**	**t**	**p**	**p_{tukey}**
Voiced	-	Devoiced	3.08	20	1.61	0.124	0.266
Voiced	-	Neutral	-1.40	20	-1.01	0.325	0.580
Devoiced	-	Neutral	-4.48	20	-2.78	0.012	0.030*

3.2.2 /t/ Variation

The mean personality rating scores for the /t/ variation component of the experiment are shown in Figure 5. On the first visual inspection, it appears the flap was associated with the highest (most positive) ratings for charisma for half of the speakers (CA, MC, MK) while the standard /t/ rates the highest for the other half of speakers (BF, CG, GR). Unlike Experiment 1, there is more of a pattern with the highest-rated variant, either the [t] or the flap, having been preferred depending on the speaker. The lowest rated (most negative) personality scores appear to have been found more consistently across speakers for the glottal stop variant.

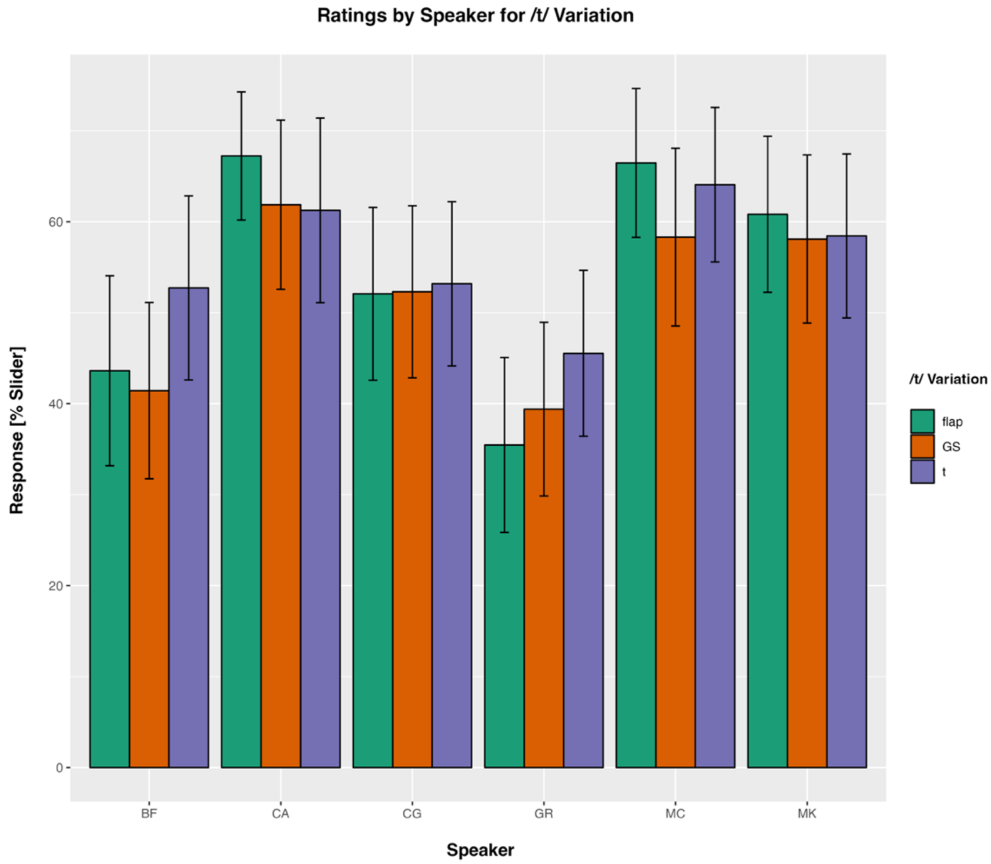


Figure 5: Personality ratings by online participants for samples from the 6 speakers' productions of the three voicing variants of /t/. The y-axis displays the slider position percentage with 0% corresponding to the left extreme value and 100% as the right extreme value of the slider. The x-axis shows the different speakers, and the coloured bars the different allophonic variation levels: flap([ɾ]), glottal stop (GS; [ʔ]), and standard t ([t]).

As in Experiment 1, a 6 x 3 repeated measures ANOVA was performed to examine how the /t/ variants affected listeners' ratings of charisma. The independent factors were speaker and /t/ variant (i.e., /t/, alveolar flap or glottal stop). Unlike in Experiment 1, in the current experiment listeners were only presented with one variant of /t/ from each speaker. For example, listener 1 would only hear the standard [t] variant from speaker 1, the flap variant from speaker 2, etc., making this a Latin square design. Because not all participants were exposed to all condition combinations, analysis was performed for triplets of three participants in the same manner as for the devoiing experiment. The results from the repeated measures ANOVA found statistical significance for speaker ($F(5, 100) = 58.61$, $p = < 0.001$, $\eta^2_p = 0.746$), /t/ variant ($F(2, 40) = 6.89$, $p = 0.003$, $\eta^2_p = 0.256$) and for the interaction between speaker and /t/ variant ($F(10, 200) = 2.18$, $p = 0.020$, $\eta^2_p = 0.098$). Tukey's post hoc difference tests for the /t/ variant condition (Table 6) indicated significantly higher ratings in the standard [t] compared to the glottal stop condition as well as higher ratings in the flap compared to the glottal stop condition. No statistical significance was found between /t/ and the flap.

Table 6. Post hoc tests of personality ratings in /t/ variant conditions for online participants.

Comparison							
/t/ Variant		**/t/ Variant**	**Mean Difference**	**df**	**t**	**p**	**p_{tukey}**
[t]	-	flap	0.833	20	0.576	0.571	0.835
[t]	-	glottal stop	4.182	20	3.488	0.002	0.006**
flap	-	glottal stop	3.349	20	3.902	< .001	0.002**

The interaction between speaker and /t/ variant was statistically significant. However, upon further inspection using Tukey's post hoc comparisons (see Table 7), no significant differences between the /t/ variants survived the Tukey correction for any of the six speakers individually. Much like in Experiment 1, there was not enough power to reveal consistent allophonic differences for individual speakers due to the high number of tests performed. However, as can be seen in the uncorrected column (the pure *p*-values without correction), similar tendencies towards lower personality ratings for glottal stops compared with the two other variants were seen for four of the six speakers.

Table 7. Post hoc results of /t/ variation effects on personality ratings separated by speaker and allophonic variant (flap [ɾ], glottal stop (GS; [ʔ]), and voiceless alveolar plosive, *i.e.,* standard [t]. Columns show the results for mean difference, t score, the uncorrected *p-value*, and the Tukey's test *p-value*.

Speaker		**Mean Difference**	***t***	***p-value* uncorrected**	***p-value* Tukey adjusted**
BF	[t] vs flap	9.58	2.29	0.033*	0.682
	[t] vs GS	13.31	3.95	< 0.001***	0.052
	flap vs GS	3.74	0.77	0.448	1.000
CA	[t] vs flap	-7.41	-2.11	0.048*	0.785
	[t] vs GS	-2.27	-0.55	0.586	1.000
	flap vs GS	5.15	1.86	0.078	0.897
CG	[t] vs flap	1.40	0.36	0.723	1.000
	[t] vs GS	1.52	0.36	0.724	1.000
	flap vs GS	0.12	0.03	0.979	1.000
GR	[t] vs flap	8.73	1.995	0.060	0.841
	[t] vs GS	6.98	2.27	0.034*	0.694

	flap vs GS	-1.75	-0.37	0.713	1.000
MC	[t] vs flap	-4.18	-1.19	0.247	0.998
	[t] vs GS	7.12	1.97	0.063	0.854
	flap vs GS	11.30	3.07	0.006**	0.255
MK	[t] vs flap	-3.12	-0.94	0.360	1.000
	[t] vs GS	-1.58	-0.41	0.688	1.000
	flap vs GS	1.54	0.48	0.639	1.000

3.3 Discussion and Conclusions

The two experiments investigated the possibility that allophonic variation in speech affects listeners' impressions of charisma. Allophonic effects on personality ratings were studied in two experimental contexts: in Experiment 1, participants listening to speech excerpts in person in a laboratory setting and in Experiment 2, listeners participating online.

Our results demonstrate some systematic effects of allophones resembling those reported by Niebuhr (2017), who examined the reduction in prosodic and segmental domains and found systematic negative effects for reductions, with productions having more reductions being rated as less charismatic. Like in Experiment 1, in Experiment 2 we found a strong effect for speaker as demonstrated in the interaction plots (Appendix D, Figure D1 and Appendix E, Figure E1). Despite this strong effect, there was a significant effect found for some of the phonetic variables.

Negative effects of final devoicing were significant in both Experiment 1 and 2. Negative effects of allophonic variation were also found for glottal stops in the /t/ variation component for both experiments.

This effect for phonetic variables appears to be missing for the final consonant devoicing component, at least for the research question we are interested in, i.e. the difference between final voiced versus devoiced items (please recall that only neutral versus devoiced was significant, with devoiced ratings being lower (negative) in all personality traits compared to neutral environments). Because of this outcome, the lack of statistical significance between voiced versus devoiced and voiced versus neutral cannot provide any useful information regarding the overall influence of devoiced versus voiced final consonants. Since there is no difference between voiced and devoiced conditions nor voiced and neutral conditions, the result of statistical significance between devoiced and neutral is surprising especially with the small statistical power. In any case, our results make us reject our first hypothesis: for online participants final consonant devoicing does not significantly increase ratings of personality and therefore does not positively increase perceptions of charisma. Despite the non-significant results found for the main effects of allophone variation, speaker effects were statistically significant suggesting ratings of charisma appear to be more speaker-dependent rather than variant-dependent.

For the /t/ variation component, we found no significant differences between [t] and the alveolar flap condition. This is interesting as [t] is the underlying representation of the allophonic variation, but in standard North American English, the [ɾ] is the most common allophone produced approximately 90% of the time, followed by the [t] at approximately 6%, and then the glottal stop with significantly reduced frequency of usage around 4% (Patterson & Connine, 2001)[16]. For listeners, it appears the [t] and [ɾ] variants are found on the same level and can be used interchangeably without influencing perception and thus the ratings of charisma in a speaker. Whereas the [ʔ] is clearly different and not considered as a standard variant for producing the underlying /t/ (for North American English). This is why listeners consistently rate the glottal stop as lower than the more acceptable variants of [t] and [ɾ]. As there are no differences between [t] and [ɾ] (no statistical significance), it appears to not matter to listeners if the underlying phoneme distribution manifests as the full allophone of [t] or the reduced allophone [ɾ], at least when rating the charisma of speakers.

Comparisons between the [t] and [ʔ] showed statistical significance with [t] rating higher (more positive) in personality trait ratings. The comparison between [ɾ] and [ʔ] also found statistical significance with the [ɾ] ratings higher (more positive) in personality traits. At this juncture, an explanation for these results in terms of underlying representation versus frequency remains unclear. Further research would need to investigate this to obtain a clearer comprehension of each domain's impact. This may be due in part to how clearly the different variants could be perceived in the online experiment compared to the in-person experiment which has controlled audio equipment.

Despite an initially statistically significant interaction between speaker and /t/ variant further inspection found no significant pairwise Tukey-corrected differences between the /t/ variants were found for any of the six speakers in the online experiment (and only one effect for one speaker in the in-person experiment). This is in part due to insufficient power because of the high number of tests performed to reveal consistent allophonic differences for individual speakers and the smaller number of trials within speaker. As mentioned, the uncorrected *p-value* scores without correction allude to similar tendencies towards lower personality ratings for glottal stops compared with the two other variants were seen for four of the six speakers.

Cumulatively, these results show the choice between [t] and [ɾ] does not appear to affect personality trait ratings to the extent of detecting statistical significance in participants samples of around 20 and therefore did not reveal reliable effects on ratings of charisma. In contrast, the glottal stop was most detrimental to ratings of personality traits having the lowest (most negative) ratings fr the tested population. This suggests that avoiding the glottal stop and opting for either [t] or [ɾ] pronunciations would result in the highest (most positive) ratings in personality traits. These results for the /t/ component appear to be the case at least for standard North American varieties of English. However, they might not hold true for other varieties of English. Future studies using an identical

[16] In the study by Patterson & Connine (2001) the rate of variant production differed dependent on morphological complexity of words, but generally followed similar trends of [ɾ] most produced, [t] second most produced – at a much lower rate – and [ʔ] produced the most infrequently.

experimental paradigm within different English varieties, such as standard British dialects, or the speech of non-native English speakers with different underlying representations in their native language phonology, like German or Russian, may provide additional interesting results for this theory.

4. General Conclusions

When contrasting the results of Experiment 1 (in-person) and Experiment 2 (online), the overall results of the online study indicate that the production of the glottal stop in /t/ variation component influences personality trait judgements and ratings of charisma (lowers charisma scores) while final consonant devoicing does not have any effect, at least with regard to our research question (voiced versus devoiced final consonants). This is important, especially when building upon the results of Experiment 1. There appears to be some contradictory outcomes regarding allophonic variant selection in individual speakers' speech productions. While both experiments demonstrated that allophonic variation and allophonic variant selection does appear to impact ratings of charisma, the partially inconsistent detailed results of each experiment when compared against one another suggest experimental presentation matters. For in-person experiments, results suggest that generally maintaining voicing in final consonants increases ratings of charisma more positively than devoicing. Perhaps acoustic differences between voiced and devoiced in the final consonant devoicing environments are more easily perceivable especially for in-person experiments. Or the participants for the in-person are more familiar with "devoicing" as a concept, as many if not all the participants were linguistics undergraduate students, therefore more readily interpret the differences between voicing. For online participants, which were a randomized pool from Prolific, perhaps the distinction between voiced and devoiced stimulus were less salient or perceived as the same (e.g., all voiced, or all devoiced).

For the /t/ variation component, the selection of the /t/ allophonic variant does not affect charisma ratings for in-person participants, with no significance found for any of the comparisons for /t/ variants. This could have been the result of other factors such as too little power, environment of experiment presentation (in-person) or the selection of the six speakers with some presenting as highly charismatic based on other acoustic properties of their voice such as voice quality, etc. Factors such as these many have left no room for allophonic effects to be detected in-person. Online participants however showed statistically significant differences between [t] and the glottal stop (with the [t] rating higher or more positively for charisma than the glottal stop), as well as between the flap and glottal stop (with the flap rating higher for charisma). Again, the differences in results between in-person (showing no statistical significance between any of the variants) and online (showing statistical significance for the previously mentioned comparisons) suggests platform plays a role in listener (participant) perception of charisma.

Nevertheless, the results of both experiments provide some interesting future areas of exploration of in-person versus online modalities of speech presentation. With the increased usage of online platforms as a form of data collecting, it is beneficial to gain a better understanding of the differences between speech perception and production results from both online and in-person experiments. Additionally, this study also provides further evidence for segmental influences within speech on personality trait ratings and charisma perception.

As the second experiment was similar in structure to the first experiment, many of the limitations of Experiment 1 apply to this experiment as well. These include the limited number of recorded speakers (for the stimuli generation), the cross-speaker variation which is problematic when using naturally produced stimuli (see Appendix A), the potential limitation of an established relationship between the listener and the speaker (i.e., speakers being labelled as professors to the listeners participating in the perception study). Additionally, participants conducting the experiment in an unmonitored environment could potentially have had detrimental impacts as a result of outside forces such as distractions from the task. The usage of personal computers and different types and models of headphones/earbuds could also impact results in a few ways. Firstly, the quality of both the personal computer's audio interface as well as listening device[17] could impact the quality of the experimental presentation both on the Gorilla platform, as well as the quality of the audio stimuli. Particularly with the audio stimuli, there are many fine-grained acoustic features in speech which may or may not be important to the disambiguation and perceptual acuity of the stimuli presented, thus potentially affecting ratings of personality traits.

Both experiments (Experiment 1 & Experiment 2) provide a preliminary baseline for future research on allophonic variation and its impact on personality trait perception and charisma attribution. Future studies could investigate a greater selection of allophone variants, different allophonic variant categories, and a broader range of speakers with perhaps different dialectal or regional backgrounds. This additional research would further this area of study and provide greater detail and understanding of the interaction regarding allophonic variant selection/usage and its impact on personality trait perception in speakers.

5. Acknowledgements
This work was partly funded by NSERC grant RGPIN-2018-06518 to the second author.

[17] Headphones; or as worst case: using a laptop's speakers although the online listeners were instructed to use headphones, but this was not an aspect we could not monitor.

6. References

Antonakis, J., Bastardoz, N. & Jacquart, P. (2016). Charisma: an ill-defined and ill-measured gift. *Annual Review of Organizational Psychology and Organization Behaviour, 3*, 293-319.

Boersma, P. & Weenink, D. (2023). [Computer program]. Praat: doing phonetics by computer (version 6.3.10). Retrieved 3 May 2023 from http://www.praat.org/

Fischer, K., Niebuhr, O., Jensen, L. C., & Bodenhagen, L. (2019). Speech melody matters—How robots profit from using charismatic speech. *ACM Transactions on Human-Robot Interaction (THRI)*, *9*(1), 1-21.

Hairer, M. (2021) [software]. Amadeus Pro (version 2.8.8). London, United Kingdom; HairerSoft.

Michalsky, J., & Niebuhr, O. (2019). Myth busted? Challenging what we think we know about charismatic speech. *AUC PHILOLOGICA*, *2019*(2), 27-56.

McCrae, R.R. & John, O.P.(1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175-215.

Niebuhr, O. (2017). Clear Speech-Mere Speech? How segmental and prosodic speech reduction shape the impression that speakers create on listeners. In *Proceedings of the International Conference on Spoken Language Processing* (Vol. 18, pp. 894-898). International Speech Communication Association.

Norman, W.T.(1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The journal of abnormal and social psychology*, *66*(6), 574.

Patterson, D., & Connine, C. M. (2001). Variant frequency in flap production. *Phonetica*, *58*(4), 254-275.

Pearsell, S. & Pape, D. (2023) The effects of different voice qualities on the perceived personality of a speaker. *Frontiers in Communication, 7,* 298. https://doi.org/10.3389/fcomm.2022909427

Rosenberg, A. & Hirschberg, J., (2009). Charisma perception from text and speech. *Speech Communication*, *51*(7), 640-655.

Tskhay, K. O., Zhu, R., Zou, C., & Rule, N. O. (2018). Charisma in everyday life: Conceptualization and validation of the General Charisma Inventory. *Journal of personality and social psychology*, *114*(1), 131.

Walker, J.A. (2015). *Canadian English: A sociolinguistics perspective.* Routledge, 78.

7. Appendices
7.1 Appendix A

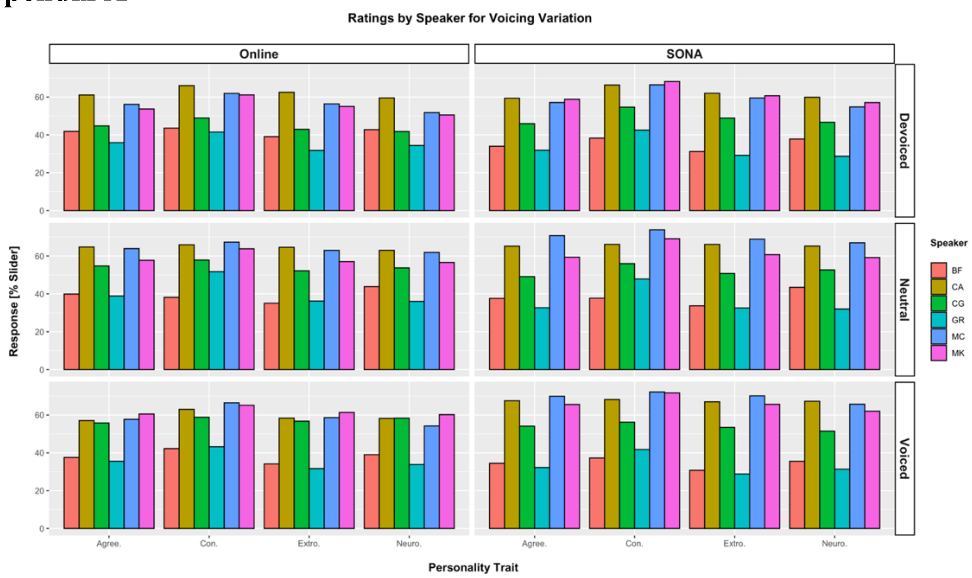


Figure A1: Speaker variation of mean scores for final consonant devoicing separated by online and SONA (in-person) participants and voicing condition (devoiced, neutral and voiced). The x-axis shows personality trait (Agree. = agreeableness, Con.= conscientiousness, Extro. = extroversion, Neuro= neuroticism). The y-axis displays the slider position percentage (0% corresponding to the left extreme value of the slider and 100% to the opposite extreme). The coloured bars represent each speaker.

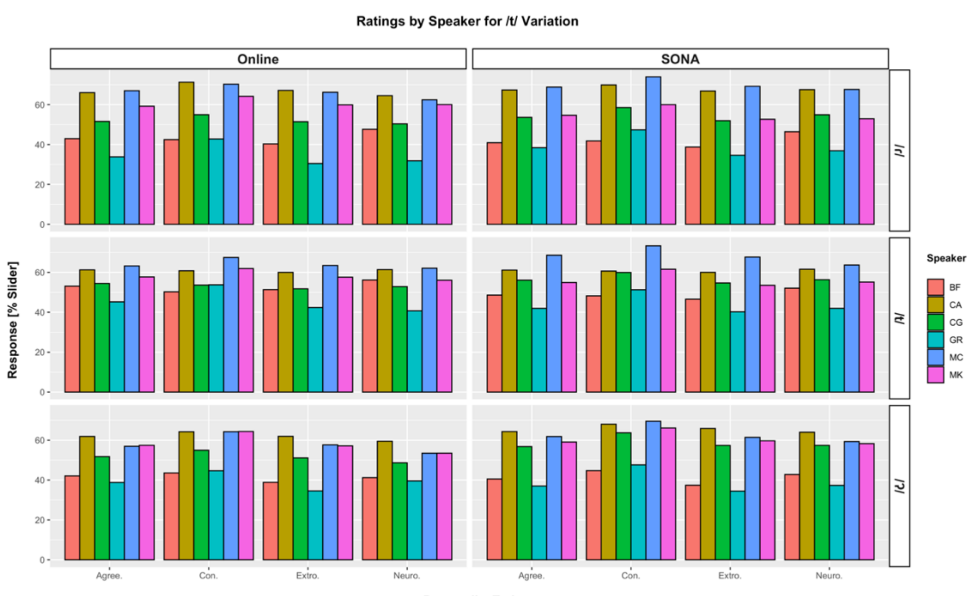


Figure A2: Speaker variation of mean scores for /t/ variation separated by online and SONA (in-person) participants and /t/ variant condition (the flap [ɾ], the voiceless alveolar plosive [t] and the glottal stop [ʔ]). The x-axis shows personality trait (Agree. = agreeableness, Con.= conscientiousness, Extro. = extroversion, Neuro= neuroticism). The y-axis displays the slider position percentage (0% corresponding to the left extreme value of the slider and 100% to the opposite extreme). The coloured bars represent each speaker.

7.2 Appendix B

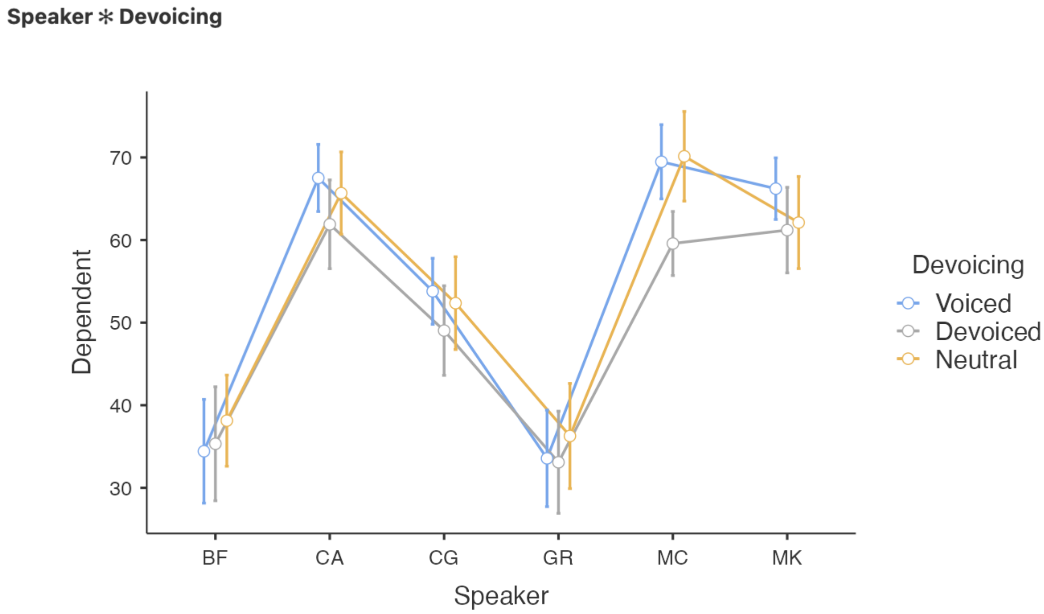


Figure B1: Interaction plot of speakers for devoicing variant for in-person participants.

7.3 Appendix C

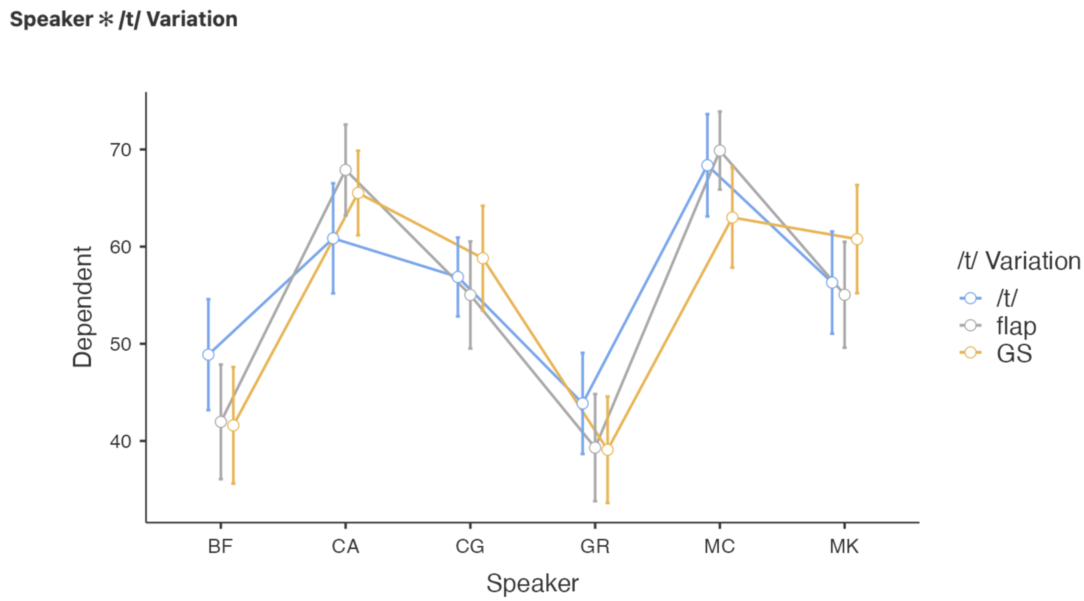


Figure C1: Interaction plot of speakers for /t/ variant for in-person participants. The variants are flap([ɾ]), glottal stop (GS; [ʔ]), and t ([t]).

7.4 Appendix D

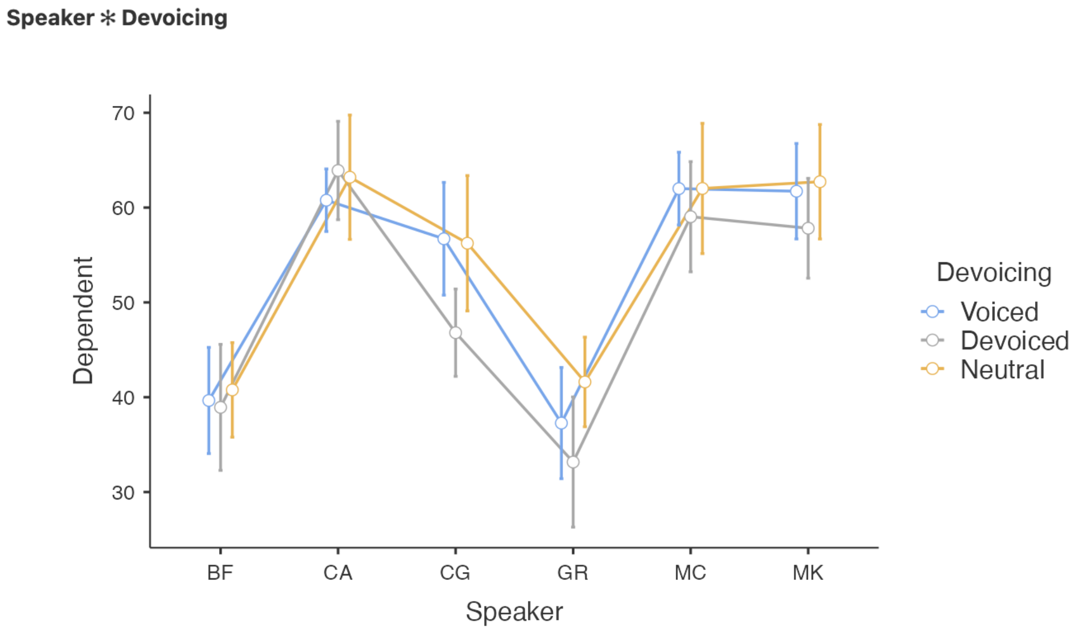


Figure D1: Interaction plot of speakers for devoicing variant for online participants.

Speaker		**Mean Difference**	***t***	***p-value* uncorrected**	***p-value* Tukey adjusted**
BF	Voiced vs devoiced	0.724	0.160	p = 0.874	1.000
	Voiced vs neutral	-1.115	-0.312	p = 0.758	1.000
	Devoiced vs neutral	-1.830	-0.585	p = 0.565	1.000
CA	Voiced vs devoiced	-3.134	-0.960	p = 0.348	1.000
	Voiced vs neutral	-2.426	-0.696	p = 0.494	1.000
	Devoiced vs neutral	0.708	0.151	p = 0.881	1.000
CG	Voiced vs devoiced	9.905	3.347	p = 0.003**	0.160
	Voiced vs neutral	0.469	0.114	p = 0.910	1.000
	Devoiced vs neutral	-9.435	-2.419	p = 0.025*	0.603
GR	Voiced vs devoiced	4.108	0.882	p = 0.388	1.000
	Voiced vs neutral	-4.336	-1.119	p = 0.276	0.999
	Devoiced vs neutral	-8.445	-2.930	p = 0.008**	0.316
MC	Voiced vs devoiced	2.964	0.826	p = 0.418	1.000
	Voiced vs neutral	-0.015	-0.004	p = 0.996	1.000
	Devoiced vs neutral	-2.979	-0.639	p = 0.842	1.000
MK	Voiced vs devoiced	3.903	1.222	p = 0.236	0.997
	Voiced vs neutral	-1.001	-0.231	p = 0.819	1.000
	Devoiced vs neutral	-4.904	-1.192	p = 0.247	0.998

Table D2: The results of the repeated measures ANOVA for the devoicing component for online participants. Results are separated by speaker and allophonic variant. Columns show the results for mean difference, standard error, t score, the uncorrected p-*value* and the *p-value* adjusted with Tukey correction.

7.5 Appendix E

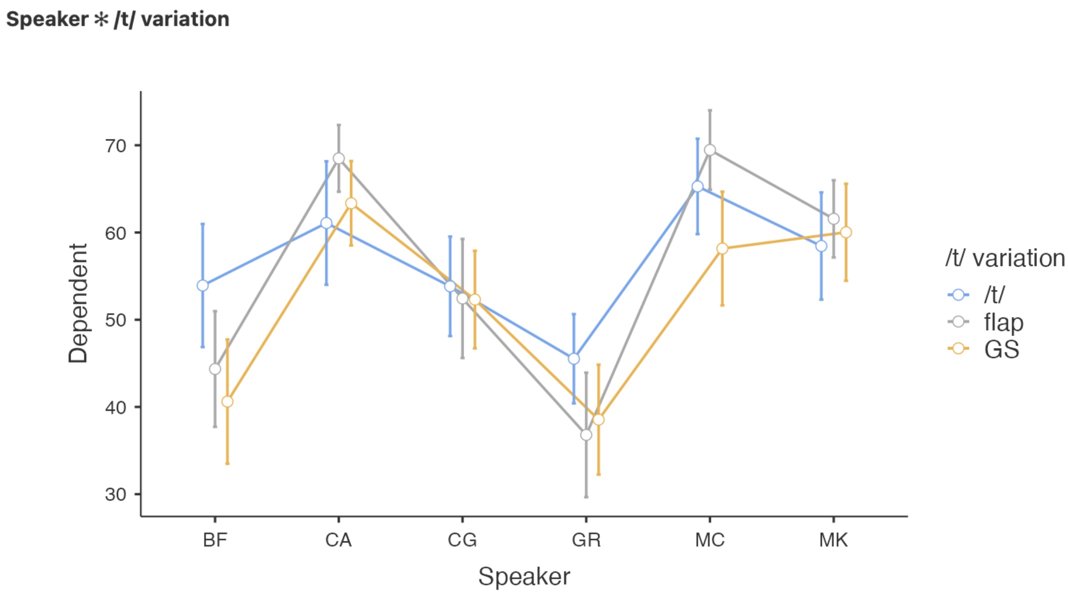


Figure E1: Interaction plot of speakers for /t/variation for online participants. The variants are flap([ɾ]), glottal stop (GS; [ʔ]), and voiceless alveolar stops ([t]).

CHAPTER 5

Conclusion & Discussion

5.1 Summary of the Thesis

This thesis investigated the effects of acoustic variation on personality trait perception, specifically charisma (and dominance), in order to answer three questions: how sound amplitude variation at different levels of language (syllable, word, sentence) affects the perceived dominance of a speaker (Chapter 2), how voice quality affects listener ratings of the personality of a speaker (Chapter 3), and how allophonic variation influences the perceived charisma of a speaker (Chapter 4).

The study of how acoustic differences in the speech signal affect the perception of personality traits, and, more specifically, how traits are linked to charisma, can offer valuable insights into the nature of perceptually relevant types of variances in the acoustic speech signal, and their implications on impressions of a given speaker character. Certain features of acoustic signals, such as voice quality or amplitude, have been reported to produce differences in perceived personality trait ascription for speakers, as discussed in Chapters 2 to 4. This thesis contributes to the existing literature by addressing variables not previously explored, such as allophonic variation, or only partially explored, like the multidimensional voice quality dimension. The outcomes of the current research provide empirical evidence for psychological consequences of perceptual acoustic variation differences. They also demonstrate their impact on personality trait attribution. Lastly, they allow the development of a more concrete definition of *charisma.*

The remainder of this chapter will include a discussion of the research questions, a summary of how each of the chapters within this thesis addresses these questions and the key findings of each chapter. It ends with a more in-depth discussion of the findings alongside the contributions these findings offer to the existing body of knowledge and future study directions.

5.2 Overview of the Research Results of the Present Thesis

5.2.1 How does amplitude variation at different levels of language affect the perceived dominance of a speaker?

Chapter 2 investigated whether loudness differences, as the perceptual correlate of acoustic amplitude variation, at different linguistic levels plays a role in dominance ratings. Chapter 2 also examined whether social and physical dominance are rated differently for these acoustic amplitude differences. To address this question, Chapter 2 split the linguistic analysis into three different hierarchical levels (syllable, word, and sentence) and dominance into two categories (social and physical dominance). Utterances produced by six speakers were presented. The amplitude of the speech was manipulated separately at each of the linguistic levels. Listeners had to answer a series of questions about social and physical dominance.

By splitting the linguistic level of analysis into syllable, word, and sentence levels, we attempted to see the effects of amplitude manipulation within each of these domains. Previous research has focused primarily on the higher linguistic levels, i.e. sentences (Scherer et al., 1973; Harrigan et al., 1989; Buller & Burgoon, 1986; Aronovitch, 1976), while lower levels remained unexamined. Investigation of word and syllable level provides insights into finer detailed perceptual discrimination effects of acoustic differences. Chapter 2 found that different linguistic levels indeed significantly impacted perceptions of dominance in speakers: The *syllable* condition with increased amplitude and *sentence* condition with increased amplitude resulted in statistically significant effects on perceptual ratings, but in different directions, with the enhanced *syllable* condition decreasing perceptions of dominance while enhanced *sentence* condition was found to increase dominance ratings. In contrast, the ratings in the reduced-loudness *syllable* condition or enhanced-loudness *word* condition, did not find statistical significance for their effects on dominance ratings. Furthermore, Chapter 2 also found no significant differences comparing the ratings given of social and physical dominance.

5.2.2 How does voice quality affect listener ratings of personality for a speaker?

To answer a question about the social perception of voice quality, Chapter 3 investigated the effects of within- and across-speaker differences of laryngeal and supralaryngeal voice quality variations on the ratings of personality traits from the *The Big 5 of Personality* model. The vocal qualities of interest in Chapter 3 were modal voice, breathy voice, creaky voice, nasality (hypernasality), and smiling. Listeners were presented an auditory stimulus consisting of speech with one of the vocal qualities of interest manipulated. Listeners then rated the speaker using ten statements and corresponding sliding scales matching to a sub-section of *The Big 5* test that had previously been shown to be linked to overall charisma perception. All listeners heard and rated all vocal qualities for all speakers.

The results of this study found each vocal quality to have statistically significant effects, meaning that all vocal qualities influenced ratings for personality traits. The results of Chapter 3 also demonstrate the different degrees of voice quality's saliency for within-speaker. Certain voice qualities, such as smiling, and creaky voice had larger effects on perceptual personality trait ratings than other qualities, like nasal or breathy voice. Results also show a trend across ratings all personality traits; traits either all increase together or decrease together. This aligns with the overall positive or negative perceptions of these voice qualities. By analyzing these attributes in conjunction with charisma, the shifts in perceptions of personality traits were found to correlate to shifts in perceptions of charisma. The higher the rating for a personality trait, such as extroversion, the higher the rating of charisma, whereas the lower the rating for a personality trait, the lower the rating of charisma.

5.2.3 How does allophonic variation influence the perceived charisma of a speaker?

Chapter 4 investigated the effects of two different forms of allophonic variation on perceptions of charisma. These two allophonic variations of interest were final consonant devoicing (FCD), and /t/ variation, in specific phonetic environments There were two implementations of this experiment, one with in-person participants, and one with online participants. Each experiment was divided into two blocks representing each allophonic variation component. Listeners heard multiple paragraphs with multiple instances of each allophonic variant and then rated the stimulus based on the ten statements used in the experiment in Chapter 3.

Chapter 4 found evidence of systematic effects of allophonic variation for FCD, and /t/ variation aligning with findings regarding segmental variation from Niebuhr (2017). The results of the first, in-person experiment for FCD found effects for charisma perception. The voiced final consonant variants (/v d z g/) increased perceptions of charisma, while voiceless variants (/f t s k/) decreased this perception. For the /t/ variation component, main effects were not found, however, there were effects found for specific speakers. This suggests that such effects may depend more on individual speakers' characteristics, like speaking style, timbre, or voice quality. Thus, while final consonant devoicing was shown to affect charisma perception, variations in /t/ articulation seemed to have more complex and variable impacts reliant on within-speaker qualities.

The results of the second, online experiment in Chapter 4 found statistically significant differences between devoiced and neutral conditions for FCD, but no differences between the relevant voiced vs devoiced conditions. Effects were found within speaker, like the effects found for the /t/ variation for the first, in-person, experiment, suggesting that effects of final consonant devoicing are, again, more dependent on individual speakers' characteristics. For the /t/ variation component, no effects were found between the [t] and alveolar flap suggesting their perceptions by listeners were very similar. However, comparison between [t] production and glottal stop, and the glottal stop and alveolar flap found significant effects. The glottal stop appears to decrease perceptions of charisma in speakers, while the usage of either [t] or the flap leads to more positive ratings of charisma. Both the [t] and the flap can be considered standard variants of /t/ in North American English whereas the glottal stop cannot.

The cumulative findings of the experiments in Chapters 2 through 4, as well as their addition to the current body of knowledge will be discussed in further detail below.

5.3 Broad Overview of Thesis Results

5.3.1 Linguistic level of perceptually relevant amplitude differences matters for dominance ratings.

Chapter 2 found that certain linguistic levels contained pertinent and observable acoustic differences for dominance ratings while other levels appeared to fall outside the listener's scope of perceptually relevant information. The results in Chapter 2 demonstrated that loudness increases in *sentence* and *syllable* linguistic levels produced meaningful perceptual differences for listeners. For the *sentence* level, enhancements of amplitude increased perceptions of dominance while increases in amplitude at the *syllable* level decreased perceptions of dominance. Interestingly, other levels, such as increases of amplitude in the *word* condition found no effect. No effects were found in reductions in amplitude specifically at the *syllable* level. These findings offer new insights into perceptually relevant domains for dominance ratings, whereas previous literature has generally focused on a singular, higher-level linguistic domain (Scherer et al., 1973; Harrigan et al., 1989; Buller & Burgoon, 1986; Aronovitch, 1976).

The results match current literature regarding prosodic mapping of amplitude and loudness in North American English. Firstly, increased dominance ratings for enhanced loudness at the *sentence* level in the experiment in Chapter 2 confirms the outcomes of previous research by Aronovitch (1976) and Tusing & Dillard (2000). They found that increases in amplitude for entire utterances increase perceptions of dominance. A novel result of the study in Chapter 2 found no effect for *word* level. This may be explained by the fact that this amplitude increase at word level is extensively used for prosodic marking of focus in English and may not be relevant for personality trait ratings due to this heavy usage for prosodic marking. This prosodic marking through amplitude variation, in principle, could also apply to the syllable level as speech utterances contain natural amplitude variation throughout production. However, only the *reduced* condition of *syllable* level found no confounding effects of prosodic cues and personality perception whereas the *enhanced syllable* condition found statistically significant results. These results contrasting results of *enhance word* level versus *enhanced syllable* level suggest amplitude increases across linguistic levels affect prosodic markings differently with some being perceptually relevant to personality trait perception (i.e., *syllable* level) while others are not (i.e., *word* level). Hence, enhanced amplitude differences at the word level remain detached from dominance and personality trait ratings and are treated as irrelevant by listeners when it comes to personality ratings.

For the syllable level, dominance ratings decrease with increasing amplitude differences for stressed syllables. This is contrary to what was seen with *sentence* level loudness increases. Previous research (Niebuhr et al., 2016; Niebuhr et al., 2020; Bosker, 2021) found that a clearer, more contrastive rhythm makes speakers sound more charismatic. Such a contrastive rhythm includes larger intensity differences between syllable edges and the nuclei as well as between stressed and unstressed syllables. The connection between charisma ratings and dominance found by Niebuhr et al. (2016; 2020) and Bosker (2021) demonstrate that increases in perception of charisma decreases perceptions of dominance. Initially, the rational for separation of dominance into types (physical and social) for the experiment in Chapter 2 was to attempt to capture features of charisma. Specifically, to investigate the effect of social dominance (defined by Muller and Mazur, 1997, p. 570) correlating to charisma. As mentioned above, previous research from (Niebuhr et al., 2016) found that increases in charisma equate to decreases in dominance, meaning that the effects found in Chapter 2 regarding increased amplitude for syllable could be a explained as a result of a clearer, more pronounced and/or contrastive rhythm. Therefore, this results in decreased perceptions of dominance (but more increased perceptions of charisma in turn) for *syllable* level, and increased perceptions of charisma (but more decreased perceptions of charisma) for the enhanced *sentence* level condition.

It is unclear why decreases in intensity differences on the syllable level did not affect dominance ratings at all. Perhaps this is tied to contrastive rhythms as well. With a flatter amplitude structure and therefore less contrastive structure between stressed and unstressed syllables, the acoustic differences which may be the relevant factors to produce perceptual judgements are no longer present. Without these natural differences, the absent differences do not reach listeners' discrimination threshold and are therefore treated as irrelevant.

5.3.2 Establishing a more concrete, empirical measurement of charisma

Within research, definitions of charisma remain vague in terms of empirical classification, if a definition is provided at all. From an investigative perspective, this leads to questions regarding what exactly a listener is judging when questioned about the presence or absence of charisma within a speech production. Chapter 2 first attempts to find a more concrete interpretation of this term through the distinction of two different types of dominance: social dominance and physical dominance. The intention behind the distinction between dominance types was that social dominance shared a definition with features with charisma. If the results of Chapter 2 found differences in ratings between the dominance types, it would provide a starting point for the empirical measurement of charisma. By employing the definitions provided by Puts et al. (2007), social dominance offers a potential correlation to the more social aspect of dominance and a potential starting point to classify charisma, while physical dominance links to the more physically associated traits of a dominant speaker. The results of Chapter 2, however, found no statistically significant difference between social and physical dominance This may be as a result of both dominance types equally affecting ratings of charisma and may both be highly correlated to ratings of charisma, or the results may suggest that social dominance is an insufficient category to capture charisma and charismatic traits.

Building from this, Chapter 3 took a different avenue to attempt a more empirical definition of charisma. Based on the previous research by Rosenberg & Hirschberg (2009) and Tskhay et al. (2018), the study in Chapter 3 opted for traits from the *Big 5 of Personality.* These traits were integrated into several statements which were linked to traits in *the Big 5* shown to be connected to charisma. Each of these statements was accompanied by a continuous scale which was used to capture increases or decreases in ratings. These increases or decreases varied depending on the positive or negative perception of a specific voice quality being rated. Based on the results of these scales, Chapter 3 found voice qualities appear to either all increase together or all decrease together. By comparing these increases/decreases for personality trait ratings and their association with charisma as previously demonstrated in the studies by Rosenberg & Hirschberg (2009) and Tskhay (2019), two conclusions can be established. Firstly, the higher, or more positive the ratings of all traits, the greater the perception and saliency of charisma; the lower, or more negative the ratings, the lower the perception of charisma. Secondly, since these increases/decreases in ratings synchronize across the different personality traits of the *Big 5* selected to be utilized in Chapter 3, not all personality traits (associated with the *Big 5)* need to be utilized in experimental designs in order to capture meaning analyses of charisma and its presence. Even the utilization of a singular trait of those investigated in Chapter 3 could be representative of measurements of charisma.

The personality trait findings of Chapter 3 were incorporated into the study in Chapter 4. Rather than offering four scores for each different personality trait variable, each personality trait was averaged amongst other scores of that personality trait. The final averages for each personality trait were then averaged, again, to create a singular score for each stimulus. This final score provided a representation of a weighted score of all the traits originally documented in Chapter 3 which was accurate to charisma ratings by itself. Thus, Chapter 4 had compound scores for each stimulus to indicate whether ratings were either positively or negatively correlated to charisma, reflecting the outcomes of Chapter 3.

5.3.3 Incorporation of suprasegmental features in the paradigms of voice quality

Previous research has generally focused on voice qualities which are predominately laryngeal, such as breathy (Pittam, 1985; Laver, 1968) or creaky (Esling, 1978; Scherer, 1979). However, by the definitions of voice quality provided in Chapter 3, supralaryngeal features are also included in aspects that define voice quality. Chapter 3 incorporates two supralaryngeal aspects (nasal voice and smiling) alongside other voice qualities traditionally exemplified (modal, creaky, breathy). Thus, the study in Chapter 3 combined modal voice qualities with the two additional vocal qualities of interest: hypernasality and smiling.

Both nasal voice and smiling voice produced statistically significant results with nasal voice having smaller effect size towards negative perceptions whilst smiling voice had much larger effects towards positive perceptions. The results in Chapter 3 regarding smiling and hypernasality are interesting, as research of both these qualities has been limited and there is no research, to our present knowledge, comparing within-speaker effects. Both results provide further understanding to which features of voice quality play a relevant role in personality perception: Smiling produced the strongest positive effects on perceived personality trait ratings amongst all the examined voice quality types, a novel finding of the present study.

5.3.4 Novel findings of allophonic variation and its impacts on charisma ratings

One novel area of investigation undertaken in this thesis is the effects of allophonic variation on perceptual ratings of charisma, in Chapter 4. A study by Niebuhr (2017) explored the impacts of speech reduction in segmental and prosodic domains and found systematic effects. In some forms, allophonic alteration also functions as phonetic reduction phenomena, oscillating between full and reduced phonemes (e.g. see flap/tap production compared to a fully produced plosive with extensive holding phase). Although not directly correlated, the research by Niebuhr provided valuable insight into potential effects allophonic variation may show in charisma ratings. The study in Chapter 4 found that this was the case, however, the relationship may be less clear than initially hypothesized.

Chapter 4 contained two studies, one conducted with in-person listeners and a second experiment conducted with online listeners. Despite having somewhat conflicting results for final consonant devoicing and /t/ variation across platforms (in-person vs. online), the results are promising for further investigation into this field. With FCD representing phonological patterns characteristic of non-native speakers or non-standard dialects, the utilization of this information is valuable for speech language pathologists (specifically for work with non-native speakers), second language learners, and all researchers interested in second language acquisition with a focus on English pronunciation. The /t/ variation component, on the other hand, establishes baseline measurements and understanding for studies of, for instance, dialectal variation (e.g., Australian English vs. Canadian English) as well as speech variables along the hyper/hypo dimension of articulation (e.g., fast speech vs. slower/clearer speech, or low-effort speech production).

5.3.5 Modality matters: in-person versus online experimental designs and results

The two experiments in Chapter 4 examined in-person and online participant outcomes for the effects of allophonic variation. Although not directly compared from a statistical perspective, this chapter revealed discrepancies in results between in-person and online results. In the first experiment, for in-person participants it was found that devoicing of final consonants for all speakers decreased perceptions of charisma, while online participants in the second experiment had varied outcomes. For the /t/ variations, in-person participants found no significant differences between variants, whereas online participants showed significant results, particularly when comparing /t/ and the glottal stop. These variations suggest that the method of data collection (in-person or online) may affect listener perceptions of charisma and do so in different ways depending on the modality of the stimulus presentation and location of testing. Overall, the studies highlight that allophonic variations can impact personality trait ratings and charisma, but the effects vary depending on the context and platform, suggesting further research is needed to understand these dynamics better.

Understanding the impacts of modality is critical in the present environment with more situations calling for online interactions, from work meetings to education to remote or hybrid working conditions. How in-person versus online speech is understood and interpreted can considerably change the effectiveness of speech as was demonstrated with the outcomes of Chapter 4. From a research perspective, this could also drastically alter researchers' understanding and perspective on online experimentation, specifically within the linguistic/speech-based research. Variability between the pools of participants between each experiment may be a contributing factor affecting the contrastive results found across each experiment.

5.3.6 Some acoustic signal differences are present in the acoustic signal and are relevant to perceptual discrimination while some remain irrelevant to perceptual discrimination

Comparing and contrasting the studies found within Chapters 2-4 provides valuable insight into the acoustic signal differences which are relevant to perceptual discrimination, and those which are treated as irrelevant to the task at hand. Some acoustic features of speech, such as voice quality demonstrated in Chapter 3, are salient for perceptual discriminations. Other features, like amplitude, have some aspects which are relevant to perception, like enhanced loudness at *sentence* level condition in Chapter

2, while other aspects of these features are resulting in non-significant perceptual results and thus most likely irrelevant to listeners, e.g., enhanced loudness at *word* level in Chapter 2.

One way to aid in distinguishing which aspects of an acoustic signal will be treated as relevant and which will remain as irrelevant to listeners is to examine other important linguistic features of a language. With a focus on English, the language of interest in the studies of this thesis, there may be some explanatory power to justify the results found within the studies of Chapters 2 to 4. For Chapter 2, the null results like those of the *word* level for amplitude variation may be a result of another feature of English: lexical stress. Lexical stress, or focus, is found mostly on the linguistic word level and serves meaningful communicative purposes for both speakers and listeners. Lexical stress is a prosodic dimension, and thus can be seen independently from speaker characteristics or phonemic/allophonic properties. Amplitude variation as a cue of lexical stress differences may, therefore, override any other potential significance of amplitude variations, like in this case perceptual speaker evaluation. Essentially, due to the highly meaningful interpretation of lexical stress for listeners for prosodic and sentence parsing purposes, any additional information, such as amplitude differences, which are embedded in the acoustic signal may be irrelevant to the experimental task at hand. Thus, rather than manifesting perceptual acoustic information, remain non-perceptual to listeners and fail to demonstrate any effects.

From a broader perspective, both laryngeal and supralaryngeal voice qualities, as well as the intrinsic link between the two, all appear to have discernible effects on personality trait perception. Segmental domains and geographical/sociophonetic differences also appear to have important effects. The results for various place of articulation differences, and segmental or geographical/sociophonetic differences suggest the large number of speech variables for investigation of perceptual differences is diverse, ranging from place of articulation to levels of linguistic organization.

For some linguistic levels, some trait rating effects were statistically significant across all rated speakers, demonstrating similar trends as those present in the results for voice quality in Chapter 3. In other cases, the effects appear to be speaker-dependent with certain speakers demonstrating specific and significant patterns for some traits, while other speakers display a different set of meaningful and significant patterns of distribution, as seen in Chapter 4. The results from Chapter 4 suggest that for some acoustic speech differences, the patterns are more complex than simply general trends and rely more heavily on speech-dependent attributes and speaker-specific effects. Some acoustic differences have more general trends (voice quality ratings both within and across speakers in Chapter 3) while some are more speaker-dependent (/t/ variation for Experiment 1 and final consonant devoicing for Experiment 2 in Chapter 4).

5.4 Limitations and Future Work

5.4.1 Limitations of Chapter 2

One limitation of the study in Chapter 2, as is also the case with Chapter 3 and Chapter 4, is the limited number of speakers used in the experimental design. Although six speakers can provide general overviews of acoustic differences and their impacts on personality trait attribution, it is important to note the shortcomings of a limited pool of speakers. This is in part due to the variations within occur naturally in the production of individual speech patterns resulting in speaker-specific speech patterns. These speaker-specific patterns in combination with a limited number of speakers may have affected the ratings and therefore the results of each experiment. For future studies, a larger number of different speakers would provide a more detailed understanding to all of the studies in Chapters 2-4. Different speakers or clusters of speakers could be presented to different participants to keep the length of experiments reasonable.

For the *syllable* level, it is not clear why ratings of dominance decreased with increasing amplitude differences; future research would need to be conducted to understand why this may be the case. As mentioned previously, some research has addressed clearer, more contrastive rhythms (which would include larger intensity differences between syllable edges and nuclei, as well as between stressed and unstressed syllables) and found that clearer contrastive rhythms increase charisma perceptions (Niebuhr et al. 2016; Niebuhr et al., 2020; Bosker, 2021). This may mean that increased charisma translates to decreased dominance. As no differences were found between dominance types (i.e., social, and physical) in the present study, it can be concluded that a more general description of dominance is satisfactory. The separation between dominance types does not provide a correlation to charisma or charismatic traits as was initially anticipated. The definition provided for social dominance by Muller and Mazur (1997, p. 570) suggested social dominance was the same or at least similar to definitions of charisma. Therefore, the syllable effect found in Chapter 2 may be a result of clearer, more pronounced and/or contrastive rhythm. This could also explain why decreasing intensity differences on the syllable level does not affect dominance ratings. This would be of particular interest to investigate in future studies.

5.4.2 Limitations of Chapter 3

Chapter 3 examined the effect of modal, breathy, creaky, nasal voice, and smiling voice qualities on personality perception. One limitation of the study is the lack of ability to control other factors of variability affecting personality trait ratings than the contrasted vocal qualities. Such factors would include possible social settings which could feed into the concept of charismatic speech and influence ratings, such as communication context (formal vs. informal), the environmental setting (academic, i.e., the professor attribution in this study vs. peer, within a group of friends), audience type (colleagues vs. interviewers), and other factors such as the intention of the communication (meant to be persuasive vs. casual conversation, or maybe interview-style). More specifically, the current study in Chapter 3 established a speaker-listener relationship (i.e., the speaker was defined as a professor) which could influence voice qualities to be perceived differently than if the study paradigm had been based on a different social relationship, for example, with the speaker as a close friend or co-worker. Future studies could benefit from the exploration of different social situations and scenarios to determine the effects of speaker role or to see whether social environment plays a role in personality trait ratings.

One component of this study, and the other studies within this thesis, which could be considered a large limitation is the selection of natural speech production rather than synthesized speech. Although synthesized speech can control for other aspects in speech which can influence personality trait ratings such as f0 and pitch (Puts et al., 2007; Rosenberg and Hirschberg, 2009; Quené et al., 2016; Berger et al., 2017), and amplitude or loudness (Novák-Tót et al., 2017), synthesized speech can also have drawbacks including artificial distortions or acoustic artifacts, intelligibility issues speech, etc. More critically, synthesized speech is often unnatural as a result additional acoustic noise which is integrated into the signal, or the lack of additional acoustic cues which are naturally integrated into human speech giving speakers their own unique voice. Naturalness, in combination with the implications of this thesis regarding personality trait ratings on *human* speakers, naturally produced speech from human speakers was the most appropriate choice for experimental design.

It is important to note that consequences for the use of human speech over synthesized productions were carefully considered during the design of stimuli. As naturalness of stimuli was imperative, the decision to use real speakers was the most appropriate. To limit the additional components of natural speech which may influence personality trait ratings, such as f0 or amplitude, speech productions were controlled as tightly as possible. This was accomplished through the employment of professional voice actors, as well as graduate level Linguistics students, to produce the various voice qualities of the experiment in Chapter 3. They were selected based on their adept skills for speech alongside their ability to produce these voice qualities on cue. Each vocal quality was explained and acoustically demonstrated to speakers before recording and throughout the recordings, speakers were continuously directed on how to produce each vocal quality. These directions included producing the vocal qualities throughout the whole produced sentences (i.e., from the start of the production until the end of the production), a comfortable and natural speech rate (not too slowly, or too fast), as well as limiting pitch variation (as flat and stable f0 as possible), and limited pitch variation (avoiding emphasis or stress). Once speakers were able to produce consistent voice qualities with the previously mentioned constraints, they were then recorded. Each paragraph was repeated and recorded three times for all vocal qualities with the best of these repetitions selected as the experimental stimulus for listeners (i.e., the repetition with the least variability in pitch and amplitude, with continuous production of vocal quality throughout the utterance, and a natural speech rate).

Prosodic differences were controlled as tightly as possible for speakers and conditions through the continuous direction from researchers during both the practice and recording sessions with auditory checks of the stimuli by the researchers. However, prosodic characteristics such as intensity differences or f0 were not artificially manipulated in order to avoid the introduction of artifacts and did possess some variation across speakers.

For future research, be a selection of different voice quality types would be interesting. The examined modal, breathy, creaky, nasal voice, and smiling voice are only a selected few of many different classifications of voice quality. It would be interesting to investigate voice qualities more similar to one another in phonation properties. It would also be interesting to incorporate different laryngeal settings other than modal voice (e.g. breathy and/creaky) linked with other supralaryngeal features, like nasality or smiling, to view the effects that non-modal voice productions and/or a combination of different vocal qualities would have on personality trait attribution. With the previous caveat on natural speech production versus synthesized speech, it would be interesting to investigate the effects of personality trait ratings for synthesized speech within the same experimental design to observe the differences between the two types of speech.

5.4.3 Limitations of Chapter 4

Chapter 4 investigated the possibility that phonetic level variation may affect perception of charisma. One of the major limitations of this chapter is the sampling of examples of allophonic variation. There are countless phonological patterns present in the majority of languages, with English being no exception. The study in Chapter 4, only focused on two allophonic environments: final consonant devoicing and /t/ variation within flapping and glottal substitution contexts. Although these two allophonic variations are commonly observed, they remain limited in their scope. The information these two phonological contexts of variations can provide is valuable, as the exploration of the role allophone selection plays for personality trait ascription is currently very understudied. However, it is important to be cautious not to make too broad statements regarding trends or outcomes for all allophonic variants based on the limited sample investigated in Chapter 4.

Future research could examine additional allophonic variants including some which may not be present in the phonemic inventory of English speakers as allophonic variants of the German /r/ at the end of syllables (producing a uvular fricative [ʁ] or the central vowel [ɐ]). It would also be interesting to examine the results for listeners of other language backgrounds to see if similar effects are found. This could include listeners who speaker English, but it is not their native language (L2 perception) as well as listeners who do not speak English at all and thus would judge the stimuli based on other factors excluding context and phonemic/allophonic details.

The results of Chapter 4 imply that some phonetic choices can influence perception of personality traits, while others may not play a significant role in shaping perceptions of charisma (and dominance). Further research is needed to explore *why* some allophonic variants affect personality trait ratings and how the raters' own speech patterns might contribute to the observed effects.

5.5 Conclusions

Previous research has demonstrated that the acoustic aspects/signals of an individual’s speech production can independently alter the perception of traits associated with that speaker. Previous studies also established that these perceptions can be both positive and negative in nature, depending on how specific aspects of speech are manipulated. This brings up an interesting question with respect to which features of personality would be relevant to examine. With the *impact of impression* being an overarching theme for personality trait perception, measurement of charisma in speakers was an area of interest and focus within the thesis. However, we asked a combination of two questions: 1) *which* acoustic feature differences are treated as salient markers of perceived personality attributes by listeners and thus are not ignored by these listeners, and 2) *how* a listener's perception of these variations' manifests in terms of ascription of charismatic traits in particular.

To summarize, the goal of this thesis was two-fold. The first goal was to investigate specific acoustic features in speech which may consistently signal personality differences. The aim was to explore *how* these differences influence ratings of specific personality traits in speakers. We examined three different features of speech based on gaps in the previous literature which included speech sound amplitude differences (Chapter 2), voice quality differences both for laryngeal and supralaryngeal features (Chapter 3), and allophonic variation (Chapter 4). The outcomes for all three studies resulted in novel information for the field of social perception of speech. Chapter 2 revealed interaction effects between acoustic speech volume and level of linguistic organization. Chapter 3 showed within-speaker effects for voice quality and that some qualities are more salient in charisma perception than others. Lastly, Chapter 4 revealed consequences of allophonic variation on charisma perception. These studies offer perspectives and tools for further studies in the future.

The second goal was to establish a more concrete and empirical form of measurement of charisma. Chapter 2 attempted to establish a starting point for a definition of charisma by utilizing definitions of dominance established in previous research by Puts et al. (2007). Puts et al. provided two categories for dominance: social and physical. However, the construct of social dominance emerging from the research displayed characteristics which appeared complementary to previous definitions of charisma; "a dominant person tells other people what to do, is respected, influential, and often a leader, while submissive or subordinate people are not influential or assertive and are usually directed by others." Although a separation of social and physical dominance appeared to be an ideal place to start an empirical investigation into the concept of charisma, the results from Chapter 2 found no statistically significant difference between ratings of perceived social and physical dominance of speakers. This result did not support the idea that social dominance as separate from physical dominance is a component of the concept of charisma. Ultimately the results did not provide direct assistance in the formation of a more concrete, empirical measurement of charisma. However, the possibility of perception of dominance and charisma being negatively correlated was raised as an explanation.

In Chapter 3, a new approach to charisma measurement was employed based on research by Hirschberg and Rosenberg (2009) and Tskhay et al. (2018). This study implemented scales from the *Big 5 of Personality*, in particular the traits of the *Big 5* established to be connected with charisma by Hirschberg and Rosenberg (2009) and Tskhay et al. (2018). The study reported in Chapter 3 found that increases/decreases in each of these investigated traits followed similar patterns. This suggests that the synchronization across ratings in these traits captures the concept of charisma. It further suggested that not all personality traits need to be investigated in experimental designs in order to prove that charisma is present.

To sum up, the goal of this thesis was to provide experimental evidence for the effects of different speech aspects on personality trait ratings, as well as provide a more concrete, operational definition of charisma. Aspects of speech sound amplitude differences at different linguistic levels (sentence, word, and syllable), speaker variation in terms of voice quality, and allophonic variation in both in-person and online experiments provided new insights and findings. These have both practical applications and novel theoretical perspectives on charisma attribution.

The results of the experiments within Chapters 2 to 4 of this thesis offer fruitful applications to applied linguistics, particularly the realm of text-to-speech and all AI applications were emotion translation, charisma, and more natural synthetic voices are sought. With understanding the impacts of amplitude increases, voice quality modulation, and allophonic variation on the perception of charisma, as well as the modality of speech presentation, AI can integrate these speech features to increase charisma in AI generated voices. As mentioned in Chapter 3, research by Fischer et al. (2019) found that AI voices, much like humans, benefit from the perception of charisma by human listeners through an array of charismatic speech techniques including the variation of tempo, and f0 range. The incorporation of additional speech features linked to increases in charisma could only aid in the expansion of perceived AI charisma and the effectiveness of AI-to-user interaction.

The results of this thesis can also apply to more public domains. Companies or universities could implement training for stronger, more influential, and more impactful speaking styles in order to produce more effective styles of communication in meetings or lectures. Even clinical domains such as speech language pathology can benefit from the results of this thesis. As voice qualities such as a creaky voice and/or breathy voice are often by-products of smaller and larger pathologies, taking the results of this thesis into account could improve speech therapies after a voice pathology is diagnosed.

These are just a few of the meaningful contributions this thesis can provide across a number of fields and domains. With the continuation of research in this area, the applications of speech production variation and its impacts on charisma perception are many, providing useful and meaningful contributions to numerous fields and domains.

References

Aronovitch, C.D., 1976. The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *The Journal of Social Ppsychology*, *99*(2), pp. 207-220.

Berger, S., Niebuhr, O. and Peters, B., 2017, March. Winning over an audience–A perception-based analysis of prosodic features of charismatic speech. In *43. Deutsche Jahrestagung für Akustik* (pp. 1454-1457).

Bosker, H.R., 2021. The contribution of amplitude modulations in speech to perceived charisma. *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*, pp.165-181.

Buller, D.B. and Burgoon, J.K., 1986. The effects of vocalics and nonverbal sensitivity on compliance: A replication and extension. *Human Communication Research*, *13*(1), pp.126-144.

Esling, J. H. 1978. Voice Quality in Edinburgh: A Sociolinguistic and Phonetic Study. *(Doctoral dissertation)*. Edinburgh United Kingdom: The University of Edinburgh.

Harrigan, J.A., Gramata, J.F., Lucic, K.S. and Margolis, C., 1989. It's how you say it: Physicians' vocal behavior. *Social Science & Medicine*, *28*(1), pp.87-92.

Laver, J.D., 1968. Voice quality and indexical information. *British Journal of Disorders of Communication*, *3*(1), pp.43-54.

Muller, U. and Mazur, A., 1997. Facial dominance in Homo sapiens as honest signaling of male quality. *Behavioral Ecology*, *8*(5), pp.569-579.

Niebuhr, O., Voße, J. and Brem, A., 2016. What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice. *Computers in Human Behavior*, *64*, pp.366-382.

Niebuhr, O., 2017. Clear Speech-Mere Speech? How segmental and prosodic speech reduction shape the impression that speakers create on listeners. In *Proceedings of the International Conference on Spoken Language Processing* (Vol. 18, pp. 894-898). International Speech Communication Association.

Niebuhr, O., Brem, A., Michalsky, J. and Neitsch, J., 2020. What makes business speakers sound charismatic? A contrastive acoustic-melodic analysis of Steve Jobs and Mark Zuckerberg. *Cadernos de Linguistica e Teoria da Literatura*, *1*(1).

Novák-Tót, E., Niebuhr, O. and Chen, A., 2017, August. A gender bias in the acoustic-melodic features of charismatic speech?. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 2248-2252). International Speech Communication Association.

Pittam, J. 1985. Voice quality: Its measurement and functional classification *(Doctoral dissertation)*. Queensland University, Brisbane, QLD, Australia.

Puts, D.A., Hodges, C.R., Cárdenas, R.A. and Gaulin, S.J., 2007. Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), pp.340-344.

Quené, H., Boomsma, G. and van Erning, R., 2016. Attractiveness of male speakers: effects of voice pitch and of speech tempo. In *Proceedings of Speech Prosody* (Vol. 8, pp. 1086-1089).
Rosenberg, A. and Hirschberg, J., 2009. Charisma perception from text and speech. *Speech communication*, *51*(7), pp.640-655.
Scherer, K.R., London, H. and Wolf, J.J., 1973. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, *7*(1), pp.31-44.
Scherer, K.R., 1979. Voice and speech correlates of perceived social influence in simulated juries. *Language and social psychology*, pp.88-120.
Tskhay, K.O., Zhu, R., Zou, C. and Rule, N.O., 2018. Charisma in everyday life: Conceptualization and validation of the General Charisma Inventory. *Journal of personality and social psychology*, *114*(1), p.131.
Tusing, K.J. and Dillard, J.P., 2000. The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence. *Human Communication Research*, *26*(1), pp.148-171.