Acoustic Analysis of Speech Production Across Sensory Conditions and Microphone Configurations

by

Xinyi ZHANG

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, "NOVEMBER 28, 2025"

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Prof. Rachel Bouserhal, Thesis supervisor
Department of Electrical Engineering, École de technologies supérieure

Prof. Ingrid Verduyckt, Thesis Co-Supervisor
École d'audiologie et d'orthophonie, Université de Montréal

Prof. Olivier Doutres, Chair, Board of Examiners
Department of Mechanical Engineering, École de technologies supérieure

Prof. Catherine Laporte, Member of the Jury
Department of Electrical Engineering, École de technologies supérieure

Prof. Meghan Clayards, External Examiner
Department of Linguistics and School of Communication Sciences and Disorders, McGill University

Prof. Andréanne Sharp, External Independent Examiner
Faculté de Médecine, Université Laval

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON "OCTOBER 28, 2025"

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

This PhD has been, quite simply, the fulfillment of a long-held dream. I had imagined myself as a scientist when I was a kid, then this imagination came with more specificity when I found my passion in speech and speech technology, though there were many periods of time when such a future seemed to me "unlikely". Yet here I find myself, surprised by the particular grace of this journey. It has been difficult, certainly, but it has also been deeply formative—not merely in terms of knowledge gained, but also in ways I had not anticipated.

My first gratitude is to my supervisor, Rachel. The timing of our encountering strikes me as remarkably fortunate. Over four years, she broadened my experience through varied projects, and I learned far more than expected while contributing to her lab's growth. I am grateful to my co-supervisor, Ingrid, who brought speech-language pathology and the essential human perspectives to my work. I also thank Catherine for the community she has built and sustained: an atmosphere of genuine intellectual excitement about speech technology, where I feel so captured by everything. I am grateful to Olivier, who, beyond his formal role as board president, helped me deepen my understanding of the physical and acoustical aspects of my research. Meghan helped me approach fluency with statistics in R with remarkable patience. Her calm and non-judgmental presence during my more anxious moments was a genuine comfort. I also thank Prof. Andréanne Sharp for graciously stepping in as my external independent examiner.

I would also like to thank my RHAD colleagues—especially Miriam and Gabriel, i.e. the SuperHearables—who understood this journey in ways few others could. I am grateful for the LATIS community—particularly Eija and Daniel, who taught me something about having patience and living without rigid definitions while pursuing genuine fulfillment. The GRAM, CRITIAS, EERS, and CIRMMT communities supported my work in practical ways, from creating the earpieces to providing access to specialized equipment. I thank Kévin, Fabien, and Hugo who improved my understanding and practical skills related to the miniature microphones and occlusion effects. Solenn and Elliot, and before them, Malahat and Danielle offered the particular kind of understanding and working together that comes from walking similar paths. Meeting Alex at McGill led to a friendship that has been both personally meaningful and

professionally supportive. Also, thanks to Morgan for the opportunity to audit the advanced statistical methods course, which proved even more valuable and exciting than I imagined. There are so many other people I want to thank: Zewen, navigating similar challenges in Toronto and enduring my Chinese/English rants for more than 14 years and still counting; Zhiyuan, Yinan, Jianing, and Sichun for showing me other kinds of lives while offering compassion during stress; Edythe, Ruth, Christian, and Michelle for guarding my faith; and Kris, my undergraduate mentor, whose training shaped my PhD approach and whose honest words I know could always count on.

I would also like to thank my proofreader, translator, IT support, “guinea pig”, “rubber duck”, and my best friend: my husband Félix (and Bunny Cassonade and Guimauve). He provided stability during uncertainty, perspective during weak moments, and the constant appreciation for my cooking and music when other things were “falling apart” in my often pessimistic mind.

There is more family I would like to thank. I thank my "Montreal family"—Darcy, Lily, June Rose and Jerry—for embracing me with warmth and making Montreal a home for me. My parents deserve more recognition I have given them in person and in voice. They gave me the freedom to study in Canada despite the considerable financial burden this created and the natural worry that comes from having their only child so far from home. Their pride in this achievement, expressed simply and directly, really put joy in my heart. I am grateful for them, for their unspoken but acted love that I truly see. I would also like to honor my uncle, my first university professor role model, who passed away suddenly at the end of 2021. I thought of him often during both the struggles and successes on my journey, and I carry the hope that he would have found this achievement worthwhile and that I have well lived his faith in me.

I would also like to thank the Fonds de recherche du Québec, the NSERC-OPSIDIAN program, ÉTS, and CIRMMT for the additional financial support I’ve received along the way. And, I thank all my participants who generously gave their time to my experiments.

Oh. . . I have so many thanks to give, and the more I think about it, the more grateful I am. My one last (written) thank-you, for all things and people that worked together, and I hope good deeds have indeed been realized.

Analyse acoustique de la production de la parole sous différentes conditions sensorielles et configurations de microphones

Xinyi ZHANG

RÉSUMÉ

La parole est façonnée non seulement par des processus cognitifs et moteurs internes, mais aussi par les conditions sensorielles externes et les outils technologiques à travers lesquels elle est produite et enregistrée. Les développements contemporains, tels que les environnements de communication de plus en plus immersifs et l'utilisation généralisée d'appareils portables intra-auriculaires (*hearables*), ont introduit de nouveaux défis et opportunités pour comprendre la parole dans des contextes qui reflètent plus fidèlement la vie quotidienne. L'objectif général de cette thèse est d'étudier comment l'intégration auditive-visuelle, les conditions d'écoute modifiées et les nouvelles méthodes d'enregistrement influencent la production de la parole et son analyse, dans le but de faire progresser les modèles théoriques et d'éclairer les applications dans les domaines de la communication et de la surveillance de la santé.

La première étude examine la base multisensorielle du contrôle de la parole en étudiant comment les caractéristiques visuelles et auditives d'une pièce affectent conjointement la régulation du niveau de la parole. À l'aide d'environnements de réalité virtuelle immersifs avec des acoustiques et des visuels variables, il est démontré que les deux modalités façonnent dynamiquement la production vocale, les informations auditives exerçant une influence plus forte, mais les informations visuelles modulant la parole plus tôt dans le temps.

La deuxième étude porte sur les effets combinés du bruit, de l'occlusion de l'oreille et de la déficience auditive sur la production de la parole. Un nouveau corpus de discours bilingue incluant l'utilisation d'appareils *hearables* a été développé, intégrant des enregistrements dans des conditions d'écoute systématiquement variées et un large éventail de seuils auditifs. Les analyses ont révélé des différences individuelles complexes dans la régulation du niveau vocal, notamment une réactivité réduite au bruit chez les participants présentant une perte auditive plus importante dans des conditions d'occlusion élevée. Ces résultats soulignent la nécessité d'adopter des approches de modélisation individualisées plutôt que collectives.

La troisième étude évalue la manière dont les nouveaux microphones intra-auriculaires et externes des appareils *hearables* capturent les mesures acoustiques de la qualité de la voix et de la fréquence fondamentale par rapport aux microphones de laboratoire standard. Les résultats indiquent des divergences systématiques entre les méthodes d'enregistrement, soulignant l'importance de développer de nouvelles normes pour l'évaluation de la voix avec les appareils *hearables*.

Ensemble, ces études élargissent notre compréhension de la manière dont la production de la parole est régulée dans des conditions environnementales, sensorielles et technologiques variées. En situant la parole dans les conditions multisensorielles et technologiques de la communication moderne, la thèse contribue aux modèles théoriques du contrôle moteur de la parole et fournit

des informations empiriques pour des applications dans la communication virtuelle, la sécurité au travail et les technologies portables.

Mots-clés: production de la parole, rétroaction auditive, acoustique de la parole, effet d'occlusion, effet Lombard, déficience auditive, intégration audiovisuelle, appareils *hearables*

Acoustic Analysis of Speech Production Across Sensory Conditions and Microphone Configurations

Xinyi ZHANG

ABSTRACT

Speech is shaped not only by internal cognitive and motor processes but also by the external sensory conditions and technological tools through which it is produced and recorded. Contemporary developments such as increasingly immersive communication environments and widespread use of in-ear wearable devices (hearable) have introduced new challenges and opportunities for understanding speech in contexts that more closely reflect everyday life. The overarching objective of this thesis is to investigate how auditory–visual integration, altered listening conditions, and novel recording methods influence speech production and its analysis, with the goal of advancing theoretical models and informing applications in communication and health monitoring.

The first study examines the multisensory basis of speech control by investigating how visual and auditory characteristics of a room jointly affect speech level regulation. Using immersive virtual reality (VR) environments with varying acoustics and visuals, it is shown that both modalities shape vocal output dynamically, with auditory information exerting a stronger influence but visual information modulating speech earlier in time.

The second study addresses the combined effects of noise, ear occlusion, and hearing impairment on speech production. A new bilingual speech corpus including the use of hearable devices was developed, incorporating recordings across systematically varied listening conditions and a wide range of hearing thresholds. Analyses revealed complex individual differences in speech level regulation, including reduced reactivity to noise in participants with greater hearing impairment under high-occlusion conditions. These findings highlight the need for individualized rather than group-level modeling approaches.

The third study evaluates how novel in-ear microphones (IEMs) and outer-ear microphones (OEMs) in hearable devices capture acoustic measures of voice quality as compared to standard laboratory microphones. Results indicate systematic discrepancies across recording methods, highlighting the importance of developing new standards for voice evaluation with hearables.

Together, these studies extend our understanding of how speech production is regulated under varied environmental, sensory, and technological constraints. By situating speech within the multisensory and technological conditions of modern communication, the thesis contributes to theoretical models of speech motor control and provides empirical insights for applications in virtual communication, occupational safety, and wearable technologies.

Keywords: speech production, auditory feedback, speech acoustics, occlusion effect, Lombard effect, hearing impairment, audio-visual integration, hearables

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	Semi-anechoic room condition, in Study 1
AC	Air conduction
AD	Alzheimer's disease
AIC	Akaike information criterion
ANSI	American National Standards Institute
Aud	Auditory condition, in Study 1
BIC	Bayesian information criterion
BC	Bone-and-tissue conduction
C	Classroom condition, in Study 1
CI	Confidence interval
CIRMMT	The Centre for Interdisciplinary Research in Music Media and Technology
CPP	Cepstral Peak Prominence
F0	Fundamental frequency
F1	First formant
F2	Second formant
FFT	Fast Fourier Transform
FE	Fixed effect
G	Gymnasium condition, in Study 1
GAMM	Generalized additive mixed-effects modeling

HIBiSCus	The Hearing-Integrated Bilingual Speech Corpus
HN	High noise condition, in Study 2
HNR	Harmonics-to-noise ratio
HO	High occlusion condition, in Study 2
HINT	Hearing in Noise Test
HPD	Hearing protection device
HRTF	Head-related transfer function
IEM	In-ear microphone
ISO	International Organization for Standardization
JASA	The Journal of the Acoustical Society of America
LN	Low noise condition, in Study 2
LO	Low occlusion condition, in Study 2
LOESS	Locally Estimated Scatterplot Smoothing
LME	Linear mixed-effects model
LRT	Likelihood ratio test
NN	No noise condition, in Study 2
OE	Open ear condition, in Study 2
OEM	Outer-ear microphone
PD	Parkinson's disease
PTA	Pure-tone average

RE	Random effect
RMS	Root mean square
RT	Reverberation time
se	Standard error
SE	Simulated open ear condition, in Study 2
SD	Standard deviation
SNR	Signal-to-noise ratio
SpEAR	The Speech-In-Ear corpus
SPL	Sound pressure level
Vis	Visual condition, in Study 1
VR	Virtual reality

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

dB	Decibel
dBA	A-weighted sound pressure level in decibel
dBHL	Decibel hearing level
Hz	Hertz
ms	Millisecond
Pa	Pascal
β	Estimated fixed-effect coefficient
σ^2	Variance

INTRODUCTION

0.1 Context and background

Speech production is a highly complex process that depends not only on motor control of the vocal tract but also on sensory feedback. The auditory feedback allows speakers[1] to monitor their own voice and make rapid, often subconscious adjustments to maintain intelligibility and meet communicative goals. Because of this tight coupling, alteration to how we hear ourselves may alter how we speak.

Extensive research on speech production has been conducted in silence, using front-of-mouth microphones and a "typical" acoustic condition. While such studies have yielded valuable theoretical insights, they do not fully capture the realities of everyday speech, which occurs in far more variable circumstances. In real life, auditory feedback can be disrupted or altered by background noise, room acoustics, and more. Understanding speech production under these modern, less predictable conditions is essential for both advancing theory and supporting practical applications.

Two major factors are now reshaping the context in which speech is produced and recorded. The first is technological. Immersive systems such as virtual reality can create situations in which the visual scene presented to a user does not match the acoustics of their physical environment, leading to an auditory–visual mismatch. Hearing protection devices (HPDs) and certain augmented or assistive listening systems can similarly alter auditory feedback while leaving visual cues unchanged. In parallel, in-ear wearable (hearable) technologies have expanded beyond the traditional microphone configuration to include in-ear microphones (IEMs) pointing inwards of the ear canal. These wearable devices, positioned in and around the ear, can monitor a wide range of physiological signals. IEMs, especially when pointed inward in

[1] Throughout this document, "speaker" refers to a person who produces speech. When referring to the audio device, the full term "loudspeaker" is used.

an occluded ear canal, offer advantages such as reduced environmental noise and the potential for continuous, long-term voice and other bio-signal monitoring. They hold promise for both communication and health applications, but more research is still needed to determine how their recordings compare with traditional methods and how best to interpret in-ear acoustic data.

The second major factor is demographic. The prevalence of hearing impairment is increasing worldwide, driven primarily by aging populations and noise exposure. The World Health Organization projects that by 2050, nearly 2.5 billion people will experience some degree of hearing impairment (Chadha, Kamenov & Cieza, 2021). In the United States, current estimates suggest that nearly one in four individuals aged 12 or older are affected, most often with mild to moderate loss (Goman & Lin, 2016). In Canada, 38% of adults (20 to 79 years old) have audiometrically measured hearing impairment (Statistics Canada, 2021). A study done with adults aged 40 to 79 years old with valid audiometric results showed that the majority of those who have hearing impairment have *unperceived* hearing impairment (Ramage-Morin, Banks, Pineault & Atrach, 2019); this makes early detection of hearing impairment difficult, and it suggests that traditional screening may have a limited reach and can benefit from a more naturalistic, everyday approach towards detection of hearing impairment. While research on hearing impairment has largely focused on perception, its effects on speech production remain less well understood. This is a critical gap, as many communication technologies and speech-based algorithms are developed using data from individuals with normal hearing, potentially limiting their effectiveness for a substantial portion of the population. The issue extends beyond occupational noise exposure to everyday technology use, such as earbuds, whose long-term impact on hearing when used at high volumes has been linked to risk of hearing impairment (Danhauer *et al.*, 2009; Byeon, 2021; Dehankar & Gaurkar, 2022).

These technological and demographic trends intersect in important ways. Both alter the auditory feedback available to speakers, sometimes in subtle ways, and both influence the acoustic signals

captured for analysis. Addressing these shifts requires extending existing speech production models to encompass new device configurations, altered sensory conditions, and diverse speaker populations. It also opens a promising avenue for public health: using speech production as a marker for hearing status. Speech is produced naturally and frequently in daily life, making it a readily available signal for analysis. Unlike perception tests, which require controlled listening tasks, speech samples can be collected unobtrusively, even in non-laboratory settings. Because hearing and speaking are closely linked, careful analysis of speech production could support earlier detection of hearing impairment, providing benefits both for individual health monitoring and for the design of more inclusive communication technologies.

0.2 Research objectives

As outlined in Section 0.1, contemporary speech production research faces new challenges arising from two major shifts: technological advances, such as immersive virtual environments and in-ear recording devices, and demographic changes, notably the increasing prevalence of hearing impairment. These factors alter the sensory feedback available to speakers and reshape the ways in which speech can be recorded, analyzed, and applied. To address these challenges, this thesis investigates speech production under conditions that better reflect real-life communication, focusing on how auditory and visual feedback, altered listening environments, and microphone placement influence speech output.

This work is organized into three studies, each targeting a different aspect of these challenges.

Study 1 examines the integration of auditory and visual information during speech production. Using virtual reality (VR), we manipulate both the acoustic properties of a room and its visual appearance (e.g., a classroom versus a gymnasium) to investigate how these cues interact in the regulation of speech level. By moving beyond static averages to analyze the temporal dynamics of speech level adjustments, this study provides new insight into multimodal mechanisms in

speech motor control and contributes to our understanding of speech production in immersive environments.

Study 2 addresses the combined effects of noise, ear occlusion, and hearing ability on speech production. To this end, we developed the Hearing-Integrated Bilingual Speech Corpus (HIBiSCus), a comprehensive database including participants with a wide range of hearing thresholds, recordings in English and French for three different speech tasks, and parallel data from multiple microphone placements. Using this corpus, we characterize how noise, ear occlusion, and individual hearing status jointly influence speech level in read speech. This study provides evidence of how altered listening conditions and hearing impairment shape speech production, while also offering a valuable resource for future research.

Study 3 investigates the implications of emerging IEM technologies for voice analysis. Traditional speech analysis relies on front-of-mouth recordings, but microphones embedded in hearables capture speech differently due to the filtering effects of the skull and ear canal occlusion. In this study, we directly compare voice quality measures (jitter, shimmer, harmonics-to-noise ratio (HNR)) and fundamental frequency (F0) (mean, variability, range) across in-ear, outer-ear, and standard microphone placements using data from Speech-In-Ear corpus (SpEAR) (Bouserhal, Bernier & Voix, 2019). The findings inform both methodological considerations for speech research and the potential of hearables for long-term health monitoring.

Together, these three studies contribute to a more comprehensive understanding of speech production in contemporary contexts. They highlight how sensory integration, altered listening conditions, and novel recording technologies influence speech, while also pointing toward practical applications in communication devices, accessibility, and health monitoring.

0.3 Structure

This thesis is organized into five chapters, followed by appendices containing supplemental materials that accompany the journal articles where applicable.

Chapter 1 provides a literature review, which also includes knowledge and practices learned that were essential for conducting the present research. The chapter begins with theoretical models of speech production and then moves to the experimental methods used to measure speech production, with a particular focus on the acoustic analysis of speech. Topics covered include sound recording and the functioning of microphones, sound pressure level (SPL), spectral analysis, F0, and voice quality. The chapter then examines external and internal sensory factors influencing speech production that are directly relevant to this work, including noise, ear occlusion, hearing impairment, room acoustics, and vision, as well as their interactions. Finally, the chapter discusses in-ear microphonic devices, outlining their unique characteristics, implications for speech research, and strategies for their use.

Chapter 2 presents the article currently in revision with The Journal of the Acoustical Society of America (JASA), which investigates the audio-visual interaction effects on speech production in immersive VR environments. Its supplementary material can be found in Appendix I. Chapter 3 presents the article submitted to JASA on the effects of noise, ear occlusion, and hearing thresholds on speech production. Chapter 4 presents the article in revision with JASA that examines the influence of microphone placement on the voice quality and F0 measurements of speech recordings. Its supplementary material can be found in Appendix II. Chapter 5 concludes the thesis by summarizing the main findings and providing recommendations for future research.

0.4 Contribution

The contributions of this doctoral research include journal articles, conference presentations, and the creation of an open-access database. This thesis incorporates three journal articles. As of the submission of the thesis, Study 3 (Chapter 4) has been published in JASA; Study 1 (Chapter 2) and Study 2 (Chapter 3) have undergone peer review and are under minor revision to address reviewers' comments.

Each of these three studies was also presented at conferences: a preliminary version of Study 1 was presented as an invited talk at the joint 186th Meeting of the Acoustical Society of America (ASA) and Acoustics Week in Canada (AWC) 2024 (Zhang, Shamei, Grond, Verduyckt & Bouserhal, 2024b); Study 2's initial concept was presented at AWC 2021 (Zhang, Verduyckt & Bouserhal, 2021), where it received the best presentation award, and part of its results were recently presented at the 7th CIRMMT-OICRM-BRAMS Student Colloquium (Zhang, Verduyckt & Bouserhal, 2025), also receiving the best presentation award; and Study 3 was presented at the 187th Meeting of the ASA (Zhang, Braga, Shamei & Bouserhal, 2024a).

The speech database developed as part of Study 2 provides a valuable resource for future research, although its use is restricted to researchers within Canada under current regulations.

In addition to the research presented in this thesis, I also worked on several other projects in the earlier years of my graduate studies. Among those, two have resulted in conference proceedings with oral presentations: one on the intelligibility of occluded Lombard speech (Zhang, Clayards & Bouserhal, 2022), presented at AWC 2022, and one on automatic speech recognition for Québécois French (Zhang, Berger, Tran & Bouserhal, 2023), presented at AWC 2023.

CHAPTER 1

LITERATURE REVIEW

1.1 Speech production modeling

Successful speech production requires fine planning, control, and coordination of different muscles and articulators with precise timing. Speech is fundamental to human communication, serving as our primary means of sharing information in daily life. In a communicative context, successful speech production also implies aspects such as adequate intelligibility and appropriate utilization of pragmatics.

To provide theoretical context for understanding speech as a dynamic process, it is useful to briefly consider the control architectures that have been proposed in speech production models. Many models have been developed, such as the Directions Into Velocities of Articulators model (Guenther, 2006) and Task Dynamics (Saltzman & Munhall, 1989). As reviewed by Parrell, Lammert, Ciccarelli & Quatieri (2019), these models typically employ feedback control, feedforward control, model predictive control, or hybrid combinations of these architectures. While a detailed analysis of these models is beyond the scope of this review, understanding the distinction between feedforward and feedback control mechanisms is relevant to the present discussion.

In feedback control, the system's output is monitored and used to maintain control, enabling stable reactions to external perturbations but becoming inaccurate when feedback signals are noisy or delayed. Feedforward control operates without monitoring output, allowing for fast execution but failing to adapt when facing unexpected perturbations. Model predictive control addresses some limitations of both approaches by using an internal model to predict the effects of inputs, functioning as a form of pseudo-feedback that is typically faster than actual feedback propagation. This predictive mechanism allows speakers to anticipate and respond to potential disruptions before they fully manifest, illustrating how speech production can be conceptualized as an adaptive, forward-looking process rather than merely reactive.

However, model predictive control relies heavily on accurate internal models and struggles with truly unpredictable perturbations, making hybrid approaches that combine it with feedback control—such as Kalman filter implementations (Houde & Nagarajan, 2011)—particularly advantageous.

The key insight from these theoretical frameworks is that speech production emerges as an inherently dynamic and adaptive process, continuously adjusting through the interplay of predictive planning and real-time correction mechanisms.

Putting into the context of speech production, the feedforward control facilitates rapid execution of our previous knowledge on producing speech, and the feedback control is required for correcting errors and maintaining effective speech in different scenarios (Cai, Yin & Zhang, 2020; Parrell *et al.*, 2019). For example, when speaking in a noisy environment, we increase our vocal effort, a phenomenon called the Lombard effect (Lane & Tranel, 1971); in a room with good voice support, we may be able to reduce voice level and remain intelligible, an important consideration for room acoustics design (Cipriano, Astolfi & Pelegrín-García, 2017). In a laboratory setting with more fine-grained precise perturbations, when auditory feedback of speech acoustic features such as formant and pitch is altered by shifting the frequency in real-time, people demonstrate compensatory behavior (e.g., Villacorta, Perkell & Guenther, 2007; Patel, Niziolek, Reilly & Guenther, 2011). These examples highlight the important role of the auditory feedback system in speech production.

A fundamental yet often overlooked aspect is that speakers hear their own voice during speech. This self-perception enables the use of real-time auditory feedback for monitoring and adjusting speech output. Self-hearing occurs through two primary pathways: air conduction (AC) and bone-and-tissue conduction (BC).

AC transmits sound waves through the air, effectively carrying frequencies across the entire audible spectrum. Hearing through AC begins at the outer ear. The outer ear is the visible, fleshy part of ear known as the pinna (or auricle). It acts like a satellite dish to capture sound waves from the environment. Its specific shape and folds help the brain determine a sound

source's location. Then, these sound waves are directed into the external auditory canal (ear canal), which then are directed inward to the tympanic membrane (eardrum). The vibration of the eardrum transmits the sound energy to the middle ear, which contains the three smallest bones in the body called the ossicles, and their movements further transfer the sound information to the inner ear, where the cochlea is (Tatham & Morton, 2007).

Within AC, there are two subtypes: direct AC and indirect AC. Direct AC refers to the sound that travels in a straight line through the air from the source to the ear canal. In the case of self-perceived speech, the source would be the speaker's mouth—this is the fastest and least altered version of the speaker's airborne speech signal that they perceive. Indirect AC, on the other hand, consists of the sound waves that leave the speaker's mouth, travel into the physical environment, reflect off the surfaces (e.g., walls, furniture, other objects), and then enter the speaker's ear canal. These reflected sounds arrive slightly later than the direct sound and are modified by the acoustic properties of the environment, creating what we perceive as reverberation.

BC, in contrast, transmits vibrations from the vocal tract directly through the skull and surrounding tissues to the inner ear, largely bypassing the outer and middle ear. Bone and tissue act as a low-pass filter, attenuating higher frequencies and transmitting components below approximately 2 kHz (Bouserhal, Falk & Voix, 2017b). Because of this filtering, the spectral content of self-perceived speech, which is a combination of AC and BC speech, is different from speech heard through AC alone. This difference becomes apparent when people perceive their recorded voice as unfamiliar, since recordings capture only the AC component, omitting the BC component.

Understanding the fundamentals of these mechanisms is essential for interpreting what happens when these auditory pathways are altered during speech production, a key focus of the present work.

In addition to auditory feedback, the theoretical speech control model cited also incorporates the somatosensory system. While it is not the focus of the present work, this inclusion is

well-founded: we rely on tactile (the sense of touch and pressure) and proprioceptive (the sense of position and movement) feedback from our articulators to monitor and adjust speech production, and substantial research has examined this process (e.g., Guenther 2006; Tremblay, Shiller & Ostry 2003). These observations underline the inherently multidimensional nature of speech production control. Beyond auditory and somatosensory feedback, other sensory modalities, such as vision, may also play a role, as explored in this thesis.

1.1.1 Auditory scene analysis

While speech production models traditionally emphasize the motor and acoustic processes involved in generating speech, it is also helpful to consider how speakers adjust to common perceptual challenges. One such challenge is Auditory Scene Analysis, where the auditory system separates a complex acoustic mixture that is received by the ears into distinct sound sources (Bregman, 1990). For example, this is relevant in noisy situations, where the Lombard effect can be seen as an adaptive behavior that helps the speaker's voice remain perceptually separate from background noise. Similarly, in a room with good voice support, the auditory system may be able to group the strong, early acoustic reflections with the direct sound. The reflections are integrated as part of the original, reinforced voice stream, rather than separate, interfering echoes.

1.2 Acoustic analysis of speech

While speech control models have been discussed above, it is essential to consider the experimental methods that enable such modeling. These models rely on the measurement and analysis of speech features through experiments. Accordingly, this section focuses on the acoustic analysis of speech. A brief overview of sound recording techniques is provided, followed by a discussion of the speech analysis features most used in the literature and in the experimental works presented later.

1.2.1 Sound recording

Reliable speech analysis depends on high-quality recordings, and overlooking fundamental recording principles can significantly affect the results. This section reviews the key principles of acoustic recording that directly affect sound measurements.

The first step in speech analysis is the recording of speech using microphones. A microphone is a sensor that converts kinetic energy from sound waves into electrical signals. Microphones vary in type, size, and directional characteristics, each of which can influence the quality of the recorded signal, making it important to select the appropriate microphone based on the specific requirements of the study.

The frequency response of a microphone indicates how accurately it records signals at different frequencies; an ideal microphone exhibits a flat frequency response across the relevant range, ensuring faithful reproduction of the original sound. The frequency range is particularly critical for speech analysis, which requires adequate coverage of fundamental frequencies (typically 80-400 Hz) and formant frequencies extending to several kilohertz. For more detailed acoustic analyses involving higher-order formants (F3 and above) that contribute to certain phonetic distinctions and voice timbre, extended frequency response into the higher kilohertz range becomes necessary.

Directionality significantly affects both the frequency response and the microphone's sensitivity to sounds from different angles. Most microphones are calibrated for on-axis response (0° incidence angle), with performance degrading at off-axis angles, particularly for higher frequencies. This directional sensitivity must be considered when positioning microphones relative to speakers during recording. Understanding how directional effects influence the recording allows researchers to determine whether these changes will meaningfully impact their specific measurements and analyses.

Other critical specifications include internal noise, which can contaminate recordings if excessive, and dynamic range, which is the span between the quietest and loudest signals the microphone

can capture without distortion. Dynamic range is crucial to prevent clipping of loud sounds while still detecting subtle, low-amplitude signals. When setting up recordings, researchers should therefore test their configuration beforehand, adjusting variables like microphone-to-mouth distance to optimize the recording range for the expected speech levels in their study. Microphone size also influences performance characteristics, with larger diaphragms typically providing better sensitivity and lower self-noise, while smaller microphones offer greater portability and less obtrusive placement options. The exact sound levels can be determined using the microphone's sensitivity or calibration factor, which converts voltage output into dB SPL. In the present research, different microphone sizes were selected based on experimental requirements: 1-inch, 1/2-inch, and 1/4-inch microphones for various recording distances and setups, and miniature microphones for specialized applications (discussed further in Section 1.4).

Microphones interface with sound cards to facilitate recording. While most computers include built-in sound cards, external devices are often used for experiments requiring multiple simultaneous recordings or real-time audio playback. Examples of high-performance external sound cards include National Instruments data acquisition systems, while more budget-friendly options include the Roland OCTA-CAPTURE and portable devices such as Zoom recorders. Input gains on sound cards can amplify the recorded signal; however, amplification can also increase the noise floor.

The recording environment further influences signal quality. Background noise, even at very low levels, and room acoustics, including reverberation, can affect recordings. Therefore, the use of a sound-treated room is recommended to ensure the highest possible fidelity of the recorded speech signal. Sound treatment typically involves acoustic absorption materials to reduce reflections and minimize reverberation time, along with isolation measures to attenuate external noise sources. While anechoic or semi-anechoic chambers provide the most controlled acoustic environment, the primary goal is achieving an adequate signal-to-noise ratio (SNR) depending on the purpose of the analysis (for example, voice quality anlysis typically requires a SNR at above 30 dB (Deliyski, Shaw, Evans & Vesselinov, 2006)), which could still be accomplished even without laboratory-grade sound treatment through careful selection of quiet

recording spaces and appropriate microphone. Understanding and quantifying these acoustic conditions requires precise measurement of SPL, which forms a fundamental aspect of speech signal analysis.

1.2.2 Sound pressure level

SPL indicates the strength of an acoustic wave and is correlated with the perception of loudness. SPL is measured on a logarithmic scale and represents the ratio of the sound pressure of a target signal to a reference sound pressure. It is expressed in units of decibel (dB). In air, the reference is typically 20 μPa, which corresponds to the approximate threshold of human hearing at 1 kHz for a young, healthy listener. Mathematically, SPL is expressed as $20 \log_{10}(p/p_0)$, where p is the measured sound pressure and p_0 is the reference pressure.

The sound pressure of a signal can be determined using the sensitivity factor of the microphone (also called the calibration factor), which converts voltage output into Pascal (Pa). A common approach is to calculate the root mean square (RMS) of the voltage over a specific time period, giving an averaged SPL for the signal. Accurate determination of absolute SPL requires knowledge of the microphone calibration factor and any applied input gain. However, if the research objective is to examine relative changes in SPL across experimental conditions, this calibration process is not necessary, as the linear scaling factor introduced by the calibration or gain does not affect comparisons between conditions.

SPL can be further specified using frequency weighting and frequency range to provide measurements relevant to the research question. In studies related to human hearing, A-weighting is commonly applied, reflecting the non-uniform sensitivity of the human auditory system across frequencies: lower frequencies are less perceptible, while higher frequencies are more prominent, and inaudible frequencies are excluded. C-weighting is another hearing-related adjustment, used for higher-level sounds, whereas Z-weighting provides an "unweighted" measurement across the human audible frequency range. Specifying the frequency range is important because sound pressure can vary across frequencies, particularly for signals that are

not spectrally uniform. These frequency-dependent variations in acoustic energy highlight the need to examine not just overall sound level, but the detailed distribution of energy across the frequency spectrum through spectral analysis.

1.2.3 Spectral analysis

The sound spectrum represents the frequency composition of a sound wave and is obtained through Fourier analysis, which transforms a signal from the time domain into the frequency domain. The most commonly used implementation is the Fast Fourier Transform (FFT), which efficiently computes the discrete Fourier transform of digitized signals.

Spectral analysis needs windowing, which means cutting the continuous signal into short frames for analysis. This is important because the FFT works as if the signal repeats, but real signals usually change over time. The window function reduces sharp edges at the frame boundaries, which helps avoid spectral leakage, which gives unwanted effects that happen when frequencies do not match the FFT bins. Common window types are rectangular (no tapering), Hamming, and Hanning, each balancing frequency detail and leakage reduction in different ways. Choosing a window size means finding a balance between time and frequency detail. A longer window gives better frequency resolution but worse time resolution, while a shorter window does the opposite. The reasoning behind it is that a longer window size uses more cycles of the signal, and this can allow the Fourier transform to distinguish frequencies that are more closely spaced. However, since a longer window is used, if the signal actually changes rapidly within the window, then the changes will be averaged out and thus losing its more fine, time-dependent resolution. This is also why depending on the nature of the sound that is analyzed (i.e., whether it is transient or stable), the window size should be chosen accordingly.

In a spectrum, the x-axis corresponds to frequency, while the y-axis represents the magnitude of each frequency component, often expressed in units such as dB. Various spectral analysis approaches exist depending on the application. Long-term average spectra compute the mean spectral characteristics over extended periods, useful for characterizing overall acoustic speech

characteristics. Short-term spectra provide instantaneous frequency content, while power spectral density estimates reveal the signal's energy distribution across frequencies. Another way to look at the spectrum is octave band analysis, which groups frequencies into bands that get wider as frequency increases, instead of showing every frequency in detail. This approach mimics aspects of human auditory processing.

Frequency-domain representations are particularly useful because much information about the signal is encoded in this domain. Peaks in the spectral envelope correspond to formant frequencies, which are critical for vowel identification, particularly the first formant (F1) and second formant (F2). Spectral peaks can also reveal characteristics of fricatives. Non-speech signals can be analyzed similarly to identify noise components. Comparisons of amplitude across frequencies allow researchers to examine differences between experimental conditions.

Spectral information can also be shown over time using a spectrogram, which shows how the frequency content changes as the signal moves forward. A spectrogram is made by applying windowed FFTs to short, often overlapping, parts of the signal. This time–frequency view is very useful in speech analysis because it shows dynamic features like formant transitions in consonant–vowel sequences, voice onset time, and prosodic patterns. The time detail of a spectrogram depends on how much the windows overlap and how far they move each step, with settings chosen to balance efficiency and the level of detail needed. The window size is again an important setting for a spectrogram because it controls the balance between time and frequency detail. This gives us two main types of spectrograms: broadband and narrowband (Styler, 2023). A broadband spectrogram uses a very short window, like the default 5 ms in the speech acoustic analysis software Praat (Boersma & Weenink, 2001). This gives it great time resolution, which means it is good for seeing quick changes in speech, like the bursts from plosives or the fast movement of formants between a consonant and a vowel. Its unique look is a series of vertical lines, where each line is one vocal fold vibration. The opposite is a narrowband spectrogram, which uses a much longer window, usually around 50 ms. This gives it high frequency resolution, so one can see the very distinct horizontal lines that represent the F0 and its harmonics. Because of this, a narrowband view is excellent for studying changes in

F0 (such as intonation), which are slower changes in speech. It is also worth noting that the best window size for a narrowband view actually depends on the F0 of the speaker's voice, since the window needs to be long enough to capture a couple of full cycles of the lowest F0.

1.2.4 Fundamental frequency

Aside from formants, one prominent peak in the spectrum represents the F0, and that is the lowest frequency component of the complex period sound produced by the source, and its integer multiples are called the harmonics. F0 correlates to the perceptual pitch. In speech, F0 shows the rate of vibration of the vocal folds and it is a critical feature in speech analysis.

While F0 can be estimated from the spectrum, it is more commonly measured using time-domain methods. One commonly used method, which is also implemented in the speech acoustic analysis software Praat (Boersma & Weenink, 2001), is the autocorrelation method. This method finds repeating patterns in the time signal by comparing it with shifted copies of itself. When the delay matches the vocal fold vibration period, a strong peak appears. F0 can then be estimated as the inverse of this delay. This approach works well for voiced speech because the nearly periodic vibrations of the vocal folds create clear peaks, but it can be less reliable when voicing is irregular or when there is a lot of noise.

When measuring F0, it is important to avoid errors such as octave halving or doubling. These errors occur when the algorithm fails to detect every cycle of the vibration or mistakenly identifies a harmonic as the F0. Thus, careful parameter settings by limiting the F0 floor and ceiling and validation against the waveform and perception are necessary to ensure accurate measurement.

1.2.5 Voice quality

Voice quality features are less frequently examined than other acoustic parameters but provide valuable insight into variability and the degree of "noise" in voicing. During phonation, the vocal folds vibrate to produce quasi-periodic sound waves, particularly in vowels and other

voiced sounds. When this vibratory pattern becomes irregular, listeners perceive changes in voice quality such as roughness, breathiness, or strain.

Jitter and shimmer measure the variability in this phonation process: jitter assesses the temporal variability of the F0 (i.e., the time between consecutive vocal fold vibrations) across time, and shimmer measures the variability in amplitude across these pulses.

There are two general categories of standardized methods for measuring jitter and shimmer. These parameters can be measured by comparing the time difference across any number of pulses or in sets of pre-determined length. Jitter and shimmer are inherently time-sensitive and dependent on basic properties of F0 or speech rate, thus different evaluation methods can be used for different languages and speaking environments. For instance, languages characterized by shorter syllables may benefit from a shorter analysis window. However, the number of F0 pulses per syllable depends not only on syllable duration but also on the speaker's F0, as a higher F0 naturally produces more pulses within the same time interval.

Another common measure of voice quality is the HNR, expressed in dB. HNR represents the proportion of harmonic sound to noise within a voiced signal. A higher HNR reflects a clearer and more harmonic voice quality, while a lower HNR indicates reduced harmonic energy or increased aperiodicity, often perceived as breathiness, hoarseness, or creakiness.

More recently, the Cepstral Peak Prominence (CPP) has gained attention as a robust voice quality measure (Patel *et al.*, 2018). CPP quantifies the prominence of the cepstral peak. The location of the peak corresponds to the fundamental frequency in the cepstral domain, which is the result of taking the inverse Fourier transform of the logarithm of the magnitude spectrum or power spectrum of a signal (Fraile & Godino-Llorente, 2014). The prominence of this peak is a measure of the signal-to-noise ratio in the cepstral domain. Unlike jitter and shimmer, which need accurate period detection, CPP can be measured even in very disordered voices with irregular vibration. This makes it especially useful for clinical voice assessment and for studying voice quality across languages. However, CPP is harder to interpret than jitter and shimmer,

since it does not directly reflect specific vocal fold behavior or clear perceptual qualities, which makes clinical use more challenging.

1.3 Sensory factors during speech production

After explaining the basics of speech production models and the acoustic analysis methods used in this work, we now look at the real-world factors that shape how people speak. This research studies how different sensory conditions influence speech, such as noise, ear occlusion, hearing impairment, room acoustics, and visual input, as well as how these factors may interact. These conditions are common in everyday communication and can change the sensory processing speakers receive, which in turn affects the control processes described in Section 1.1.

In this section, we provide a comprehensive review of the literature on background noise and hearing impairment. These two factors have been studied extensively across diverse contexts and populations, resulting in a substantial body of work that extends beyond the scope of the individual articles in this thesis. In contrast, the literature on room acoustics, visual input, audio-visual interaction, ear occlusion, and the interaction between noise, ear occlusion, and hearing impairment in speech production is more specialized and is reviewed in detail in the introductions of Articles 2 and 3, where it directly concerns the specific research questions addressed.

1.3.1 Noise

In the context of speech communication, noise refers to any undesired sound that interferes with the transmission or perception of speech information, either for the speaker or the listener. Sources of noise can be environmental (e.g., traffic, crowd chatter, machinery) or artificial (e.g., white or pink noise) and may differ in temporal variability. Noise is often characterized by its frequency content, typically determined using Fourier Transform. For instance, ideal white noise has equal energy at all frequencies, resulting in a flat spectrum, whereas pink noise

contains equal energy per octave and is perceived as more balanced to the human ear due to our logarithmic frequency perception.

The perceived loudness of noise is also shaped by the ear's frequency-dependent sensitivity, as described by equal-loudness curves. For practical purposes, A-weighted sound pressure levels (measured in A-weighted decibel (dBA)) are often used as a psychoacoustic approximation of human loudness perception (Meinke, Berger, Driscoll, Neitzel & Bright, 2022). Noise level measurement becomes particularly important when considering potential hearing damage. Occupational safety guidelines typically specify a maximum noise level with the maximum permissible exposure time that is calculable with the exchange rate; the exposure time is halved for every exchange rate level increase in noise. Around the world, the maximum noise level is either 85, 87 or 90 dBA for 8 hours of exposure, and the exchange rate varies from 3 to 5 dB (Meinke *et al.*, 2022). In research settings, it is important to know noise exposure limits to keep participants safe.

Speakers hear their own voice through AC and BC. In noisy environments, the AC pathways are compromised as ambient noise masks the self-generated speech signal. This disruption elicits the Lombard effect, an involuntary adjustment of speech to maintain audibility (Lane & Tranel, 1971). First described by Étienne Lombard in 1911 as a indicator of hearing impairment (Brumm & Zollinger, 2011), the Lombard effect is most prominently characterized by an increase in vocal intensity. Its onset and offset (once the noise stops) period do not take longer than a few seconds but with a temporal asymmetricity, with longer onset and shorter offset (Villegas, Perkins & Wilson, 2021).

The relationship between noise sound pressure level (noise level in short) and speech sound pressure level (speech level in short) is often quantified as the Lombard slope—the change in speech level (measured in dB) per 1 dB increase in noise level—which has been reported to range from 0.2 to 1 dB/dB (Bottalico, 2018). Variability in Lombard slopes across studies can be partly attributed to differences in noise type and spectral content. For example, Bottalico & Murgia (2023) found that mid-frequency noise (0.5–4 kHz) elicited greater increases in voice level than

low- or high-frequency noise. Other research also demonstrates that the Lombard effect is stronger in communicative contexts than in non-communicative ones (Garnier, Henrich & Dubois, 2010; Villegas *et al.*, 2021).

In addition to increased intensity, Lombard speech exhibits several acoustic modifications. Well-documented changes include increased fundamental frequency, shifts in spectral energy toward higher frequencies, vowel lengthening, spectral tilt reduction, and increases in the F1 frequency (Junqua, 1996). For example, Garnier & Henrich (2014) reported that exposure to broadband noise at 86 dB SPL lengthened vowels by 33–47 ms and shortened unvoiced consonants by about 10 ms, with females showing greater increases in the spectral center of gravity than males.

Notably, Lombard speech has been shown to be more intelligible in noise than speech produced in quiet (Lu & Cooke, 2009). This enhanced intelligibility has inspired practical applications, such as Lombard-inspired spectral shaping for speech recognition (Godoy, Koutsogiannaki & Stylianou, 2014), Lombard-style text-to-speech synthesis for noisy environments (Novitasari, Sakti & Nakamura, 2022), and intelligibility improvements for cochlear implant users (Hansen, Lee, Ali & Saba, 2020). Despite these advances, individual variability in Lombard responses—both in magnitude and in specific acoustic adjustments—remains insufficiently studied.

1.3.2 Hearing impairment

Hearing impairment represents a chronic alteration to self-auditory feedback, affecting both AC and BC pathways across various frequency ranges. Unlike temporary perturbations from noise exposure or ear occlusion, hearing impairment produces long-term changes in how speakers monitor and control their own voice.

Hearing impairment is typically classified into three main types based on the anatomical site of dysfunction. Conductive hearing impairment arises from dysfunction in the outer or middle ear that impedes sound transmission to the inner ear. AC hearing thresholds are elevated while

BC hearing thresholds remain relatively unaffected. Sensorineural hearing impairment results from damage to the cochlea or auditory nerve, elevating both AC and BC thresholds and often creating frequency-specific deficits. Lastly, mixed hearing impairment combines elements of both conductive and sensorineural loss (Dalebout, 2009).

Screening and diagnosis of hearing impairment is most commonly performed via pure-tone audiometry, in which the quietest audible sound is measured at various frequencies. AC audiometry is a more routinely performed test than BC audiometry, which directly assesses cochlear function while bypassing the outer and middle ear. Results of audiometry are expressed in decibel hearing level (dBHL), a scale referenced to normative hearing thresholds (at 0 dBHL). Thresholds up to 15 dBHL are considered normal, with higher values indicating hearing impairment (Clark, 1981). A common summary metric is the pure-tone average (PTA), calculated as the mean threshold (in dBHL) at measured frequencies for each ear. 500, 1000, and 2000 Hz are some typical frequencies to use because they cover the range most important for speech intelligibility. PTA is widely used to classify hearing impairment severity (e.g., mild, moderate, severe).

In addition to pure-tone audiometry, tympanometry is a complementary, objective test used to evaluate the function of the middle ear and eustachian tube, whose results are presented as a tympanogram. This test measures the mobility of the eardrum as air pressure is varied in the ear canal and helps diagnose conditions of the middle ear system.

Hearing impairment disrupts how individuals hear themselves. Auditory feedback plays a critical role in speech production, enabling speakers to monitor articulation accuracy and adjust vocal output to optimize the speech-to-noise ratio. Research on postlingually hearing-impaired speakers (a hearing impairment that is developed after language acquisition) in quiet has shown that their speech often deviates from that of normal-hearing speakers (Cowie & Douglas-Cowie, 1992; Lane & Webster, 1991). Historically, changes in speech production under noise have also been used as an indicator for hearing impairment, showing the potential of using speech production as an indicator of auditory health.

Several studies have explored whether specific acoustic features could serve as vocal biomarkers of hearing impairment. For example, Pittman, Daliri & Meadows (2018) examined the production of the English voiceless fricatives /s/ and /ʃ/ in children and adults with mild-to-moderate hearing impairment. In children, the distance between the spectral centers of gravity for these fricatives correlated with the degree of hearing impairment; however, no such relationship was found in adults. This may reflect the relative stability of segmental features in postlingually hearing-impaired adults (Perkell, Lane, Svirsky & Webster, 1992), although other biomarkers may still exist for this population.

As summarized in Coelho, Medved & Brasolotto (2015), postlingually hearing-impaired may be associated with a range of suprasegmental and voice quality changes, including abnormal intonation, elevated F0, altered speech rate, increased nasality, loudness deviations, elevated noise-to-harmonic ratio, greater shimmer and jitter, and percepts of vocal roughness or strain. Supporting this, Nagy, Elshafei & Mahmoud (2020) reported a correlation between hyperfunctional dysphonia and undiagnosed hearing impairment, suggesting that impaired auditory feedback can affect pitch regulation.

Despite these findings, the literature remains limited. Many studies use small sample sizes, report inconsistent acoustic characteristics across researchers, and lack standardized measurement approaches. Moreover, research has focused disproportionately on severe hearing impairment and cochlear implant users, even though the American Speech-Language-Hearing Association defines hearing impairment as thresholds above 15 dBHL. Individual variability also remains underexplored, as participants are typically grouped simply by the presence or absence of hearing impairment without considering the continuous nature of hearing impairment.

Findings on vocal intensity reflect this variability. Lee (2012) reported that postlingually hearing-impaired speakers produced greater intensities than normal-hearing controls during sustained vowel phonation in quiet. In contrast, Sørensen, Lunner & MacDonald (2024) observed no significant differences in spontaneous conversational speech. Comparable task-dependent patterns have also been documented for F0 and voice quality (Coelho *et al.*, 2015; Di Stadio *et al.*,

2025), highlighting the need for more systematic research on how these long-term alterations in self-feedback affect speech motor control.

1.4 Microphonic in-ear devices

Hearables, wearable devices located in and around the ear, provide access to a wide range of physiological and behavioral signals, including heart rate, breathing, swallowing, eye movement, and speech (Mehrban, Voix & Bouserhal, 2024; Chabot, Bouserhal, Cardinal & Voix, 2021; Röddiger *et al.*, 2022; Goverdovsky *et al.*, 2017). Many hearables incorporate miniature microphones, similar to those used in hearing aids. These microphones may be positioned inside the ear canal, oriented inward, and externally on the device, oriented outward. Despite their small size, they differ in frequency response characteristics: some maintain a mostly flat response with a decreased response at high-frequency (beyond 10 kHz), while others may feature a decreased response in the low frequency region (until 250 Hz), which can help mitigate the occlusion effect and prevent low-frequency clipping. While technical specifications of individual microphones are available, their integration within hearable devices introduces additional complexities. Understanding how to select, record with, interpret data from these microphones, and how to mitigate the unwanted differences from conventional setups has often been knowledge acquired through practical experience.

Microphone placement within the hearable strongly influences the recorded signal. If positioned at the end of a tube-like structure, resonances occur at frequencies determined by the tube's length and diameter. Proximity to the tube opening reduces this effect, which can be preferable because resonances in the speech-relevant range (e.g., around 4 kHz) can affect frequency-domain features. Such distortions can be corrected through reverse filtering, typically by recording a reference signal using a small and less obtrusive microphone (ideally 1/4-inch or 1/2-inch) with a known flat frequency response in a reverberation chamber and comparing it to the hearable microphone's recording. Using broadband excitation (e.g., white noise > 80 dB SPL) allows estimation of the transfer function, which can then be applied for both frequency and sensitivity calibration.

In-ear microphones, when placed in a sealed ear canal, capture speech with reduced environmental noise. The captured signal is influenced by both AC within the occluded space and bone-and-tissue vibrations, which cause the enclosed air pocket to vibrate. This mechanism differs from that of a dedicated BC microphone but produces a related signal. However, BC functions as a low-pass filter, attenuating higher-frequency speech information—an important consideration since certain formants (e.g., F2 for female speakers in high vowels) can exceed 2 kHz. Low-frequency enhancement, on the other hand, can facilitate the detection of non-speech physiological signals such as heartbeat.

CHAPTER 2

THE TEMPORAL EFFECTS OF AUDITORY AND VISUAL IMMERSION ON SPEECH LEVEL IN VIRTUAL ENVIRONMENTS

Xinyi Zhang[1,2] , Arian Shamei[1,2] , Florian Grond[3] , Ingrid Verduyckt[4] , Rachel Bouserhal[1,2]

[1] Department of Electrical Engineering, École de technologie supérieure, Montréal, Québec H3C 1K3 Canada

[2] Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Québec H3A 1E3 Canada

[3] Department of Design and Computation Arts, Concordia University, Montréal, Québec H3G 1M8 Canada

[4] École d'orthophonie et d'audiologie, Université de Montréal, Montréal, Québec H3N 1X7 Canada

Paper submitted for publication in "The Journal of the Acoustical Society of America" on June 10, 2025.

2.1 Abstract

Speech takes place in physical environments with visual and acoustic properties, yet how these elements and their interaction influence speech production is not fully understood. While room appearance can suggest its acoustics, it is unclear whether people adjust their speech based on this visual information. Previous research shows that higher reverberation leads to reduced speech level, but how auditory and visual information interact in this process remains limited. This study examined how audiovisual information affects speech level by immersing participants in virtual environments with varying reverberation and room visuals (hemi-anechoic room, classroom, gymnasium) while completing speech tasks. Speech level was analyzed using generalized additive mixed-effects modeling (GAMM) to assess temporal changes during utterances across conditions. Results showed that visual information significantly influenced speech level, though not strictly in line with expected acoustics or perceived room size; auditory information had a stronger overall effect than visual information. Visual information had an earlier influence that diminished over time, whereas the auditory effect increased and

plateaued. These findings contribute to the understanding of multisensory integration in speech control and have implications in enhancing vocal performance and supporting more naturalistic communication in virtual environments.

2.2 Introduction

Speech is fundamental to human communication, serving as our primary means of sharing information in daily life. We communicate through speech typically within a natural environment. The acoustic characteristics of our surroundings—whether we are speaking in an empty classroom, a crowded restaurant, or a movie theater—significantly influence how our speech reaches our listeners and, consequently, how we produce our speech to maintain effective communication.

To ensure effective communication, speakers regulate aspects of their speech production, with speech sound pressure level (speech level for short) being a particularly crucial aspect. Appropriate speech level provides a sufficient SNR for intelligibility while also maintaining a comfortable experience for both the listeners and the speakers. In addition, speech level carries additional meaning in communication, as, for example, what constitutes an appropriate level varies across different social contexts.

When we speak, we can naturally produce speech at a certain level, often without consciously thinking about it. This reflects our learned knowledge of how much effort to use to achieve a level that would sound appropriate—a mechanism known as the feedforward control. In addition, we adjust our speech level based on how we hear ourselves as we speak. We hear ourselves through two pathways: air conduction and bone-and-tissue conduction. For air conduction, it further divides into direct air conduction and indirect air conduction. This mechanism is a feedback control loop. This study focuses on the auditory feedback control; however, it is important to highlight the multimodal nature of speech production control. Notably, sensory feedback contributes to speech production control in general Guenther (2006), and perilaryngeal vibration feedback specifically affects speech level control Brajot, Nguyen, DiGiovanni & Gracco (2018). The feedforward control uses internal models to make anticipatory adjustments (Parrell

et al., 2019), such as speaking louder to someone far away; the feedback control allows real-time adjustments based on our sensory inputs (Parrell *et al.*, 2019). The Lombard effect is a well-studied example of the auditory feedback mechanism, and it influences the direct air conduction pathway. Discovered by Étienne Lombard in 1911, it is a phenomenon that describes the automatic increase of vocal intensity, along with other modifications of speech production such as increased fundamental frequency and duration. It increases the audibility of one's own voice and its intelligibility in communication in noisy environments (Brumm & Zollinger, 2011), beginning at background noise levels of 43.3 dBA (Bottalico, Passione, Graetzer & Hunter, 2017).

The interplay between these two mechanisms becomes particularly interesting when we consider that we, in fact, integrate both auditory and visual environmental information when producing speech. For example, visual information about the environment may trigger learned associations, leading to predictive feedforward control of our speech level, while auditory feedback allows us to monitor our speech level in real-time and adjust to the acoustic properties of a particular space. Despite the clear multi-modal nature of this process, research has predominantly focused on auditory influences, leaving the role of visual environmental information rather unexplored.

Recent advancements in the field of Virtual Reality have increased the use of VR devices, and this has introduced new challenges for speech production by creating environments where the auditory and visual information may not always align. In a typical VR setting, the virtual environment may present a visual context, while the actual acoustics of the environment where the user is may not match this visual virtual room; this creates a mismatch between the auditory and visual sensory inputs. In addition to VR, other factors, such as HPDs, also alter the auditory feedback and lead to the audio-visual mismatch. The effects of the interaction between the auditory and visual sensory inputs and especially when they are mismatched, remain under-explored for the process of speech production.

The current study investigates how auditory and visual information about the environment are used, integrated, and potentially interact with each other during speech production, focusing

on their influence on speech level regulation. Specifically, we look at the auditory information about the acoustics of a room and the visual information about the type of a room (e.g., a classroom versus a gymnasium). Investigating how people integrate auditory and visual sensory information to control speech level not only enhances our understanding of speech production in virtually immersive environments but also provides valuable theoretical insights into the feedback and feedforward mechanisms underlying speech motor control.

To better understand how auditory and visual information influences speech production, below we review how room acoustics affect speech level, how visual information may assist in interpreting a room's acoustics, and the existing research on the influence of the interaction between auditory and visual information in speech and its production.

2.2.1 Room acoustics

Room acoustics play an important role in how sound is delivered to our ears. One way to describe a room's acoustics is with reverberation. Reverberation characterizes the persistence of sound energy after the sound source has stopped in an environment, which is heard as a decaying sound over time. It occurs due to reflections of sound waves off surfaces in the environment, and they linger in the environment until the surfaces absorb all the energy or the energy dissipates from the distance traveled (Tsang, 2023). As summarized in Tsang (2023), it has been established that in general, reverberation increases with the size of the space and decreases with the level of surface absorption.

How we hear ourselves through the air-conduction pathways serves as external auditory feedback. Research has shown that, in addition to the Lombard effect mentioned above, the acoustics of the environment influence how one speaks. This happens because different room acoustics change how sound reflects, affecting the indirect air conduction pathway for auditory feedback. This is especially important for people who rely on their voice as part of their profession, such as teachers. They need to maintain the intelligibility and clarity of their speech while having vocal comfort, which implies a control of vocal effort (Bottalico, Graetzer & Hunter, 2016). For

example, Black (1950) found that people read louder in an acoustically dead room than in a more reverberant room, especially when the size is large. By having participants read in rooms with distinct reverberation levels plus the placement of a reflexive panel to further modify the reverberation, Bottalico *et al.* (2016) found that people decreased their speech level with the increase of reverberation.

Sierra-Polanco, Cantor-Cutiva, Hunter & Bottalico (2021) investigated the differences in speech production in artificial acoustical settings that varied in reverberation time, presence of background noise, and level of gain in the playback of one's own voice. In this study, they found that the decrease in speech level from an increase in reverberation is more pronounced in the presence of noise, with an estimated 0.22 dBA decrease between Low and High reverberation and an estimated 0.17 dBA decrease between Low and Medium reverberation. Without noise presence, the estimated difference between Low and Medium is -0.08 dBA ($p = .162$) and no difference between Low and High. The effect of gain, on the other hand, was stronger. With 5 dB and 10 dB of gain, people on average decreased their speech level by 0.31 dBA and 1.26 dBA respectively. This relates to what has been reported by Pelegrín-García, Fuentes-Mendizábal, Brunskog & Jeong (2011b). In different artificial room acoustics conditions, participants were asked to match their own voice level to a constant reference. Reverberation time, room gain, and voice support were manipulated. Both room gain and voice support show the relationship between the direct airborne energy and the indirect airborne energy. Room gain, defined in Equation 2.1, quantifies the gain that a room applies to one's own voice to their own ears, while voice support, Equation 2.2, quantifies the difference in level between one's own voice heard through direct versus indirect AC. Participants adjusted their voice level more in reaction to the change in room gain and voice support than to reverberation; in fact, it was reported by the authors that participants' speech level change concerning reverberation time was "homogeneous".

$$\text{RoomGain} = 10\log\left(\frac{\text{Energy}_{\text{Direct}} + \text{Energy}_{\text{Indirect}}}{\text{Energy}_{\text{Direct}}}\right) \tag{2.1}$$

$$\text{VoiceSupport} = 10\log\left(\frac{\text{Energy}_{\text{Indirect}}}{\text{Energy}_{\text{Direct}}}\right) \tag{2.2}$$

By the definition of room gain and voice support, it seems that there should be some correlation between them and the reverberation time, at least in a naturalistic environment, as they all describe the reverberation level of an environment. However, since the underlying hypothesis for speech modification in different room acoustics conditions is due to the auditory feedback mechanism, it is understandable that reverberation level measured by reverberation time reflects what we hear less directly than by energy ratio. The distinction between these two different perspectives becomes especially important when an "artificial" gain in the room is applied, for example, in Sierra-Polanco *et al.* (2021). In this case, the "gain" of the room leads to sidetone regulation, a phenomenon that describes the observation that people modify their voice level using the auditory feedback of their own voice. It is established that in general, sidetone amplification tends to lead people to speak quieter (e.g., Tomassi *et al.*, 2023; Siegel & Pick, 1974). Indeed, it is a phenomenon that is closely related to the Lombard effect, but it is much less studied.

2.2.2 Visual information of a room

Vision is an important sense of ours and research has shown that people use visual information in the process of speech production. For example, it is well established that people increase their voice level when speaking to a listener at a further distance (e.g., Pelegrín-García, Smits, Brunskog & Jeong, 2011a; Bouserhal, Bockstael, MacDonald, Falk & Voix, 2017a), and the information of distance is easily extracted and estimated through vision.

As summarized in Tsang (2023), vision also importantly contributes to spatial perception (e.g., the size, shape, and volume of a room) and material perception (e.g., fabric, wood, metal, etc.). These two aspects of the environment play an important role in determining the reverberation of an environment, and the visual information associated with them is effortless to extract. However, whether people indeed use visual information to form an expectation about a space's reverberation remains unclear.

Tsang & Mannion (2022) showed participants simulated audio-visual stimuli, where they were presented with a visual environment first and then a speech utterance was played whose characteristics either matched or did not match with the visual environment's reverberation. Participants answered whether they thought the audio they heard was recorded in the visual environment they saw. They found that participants were more likely to respond correctly when the audio and visual conditions matched, and participants were more likely to falsely report a match when the visual environment they were presented with had similar reverberation times as the actual environment. These results show that people can use visual information to form an expectation about the acoustic environment. This result aligns with the findings in, for example, Defays *et al.* (2015), arguing that visual information does play a role. However, this is still a topic of debate, as there exist studies such as Schutte, Ewert & Wiegrebe (2019) which showed that reverberation perception was not influenced by visual information in virtual environments.

2.2.3 Interaction between environmental auditory and visual information

Research on the interaction between environmental auditory and visual information related to speech is sparse. A few studies, including those by Tsang (2023), Ahrens & Lund (2022), and Daşdöğen *et al.* (2023), have explored aspects of this topic, but only Daşdöğen *et al.* (2023) has examined speech production directly.

Daşdöğen *et al.* (2023) measured vocal sound pressure level (SPL) across two acoustic conditions (anechoic and reverberant) and two room sizes (small and large), in addition to two baseline conditions designed to isolate single modalities. Visual deprivation was achieved through wearing a black sleep mask, while auditory deprivation involved closing off participants' open-back headphones. This gives in total nine audio-visual conditions.

The findings indicated that speakers produced higher vocal SPL in less reverberant environments and in larger spaces. Although the authors reported interaction effects between acoustic and visual conditions, they did not provide explicit statistical testing of interaction terms. Furthermore,

examination of their SPL data across conditions suggests additive rather than truly interactive effects between modalities.

Several limitations constrain the interpretation of these findings. First, the study used non-spontaneous speech (counting, sustained vowels, and short read sentences) rather than spontaneous communication. Participants spoke with no actual listener present, even though they were instructed to "speak so that everyone can understand you" to encourage communicative intent. Second, the visual manipulation confounded room size with speaker-listener distance, as no constant speaker-listener distance was maintained across conditions. This confound prevents clear attribution of vocal adjustments to room *size* versus proximity to the intended communication partner. Third, the auditory baseline condition also presents methodological concerns, as it fundamentally alters participants' auditory self-monitoring process and likely increased bone-and-tissue conducted speech perception, potentially changing the baseline vocal SPL. Lastly, vocal intensity was assessed using utterance-level averages, missing the potential temporal dynamics in vocal adaptation that may be crucial for understanding real-time environmental adjustments.

In the current study, we aim to build on these findings and limitations by examining how people dynamically adjust their speech level over time within different visual and auditory environments, with particular attention to interactions between these sensory elements. To our knowledge, this is the first study that goes beyond static averages to investigate the temporal dynamics of speech level modulation in such settings. We address the following research questions:

- Does reverberation indeed influence speech level?
- Does visual immersion impact speech level?
- Are auditory and visual information integrated in controlling speech level? If so, any interaction?

2.3 Methodology

2.3.1 Participants

Twenty fluent French speakers participated in this experiment (F: 13; M: 7). The exclusion criteria were that they should not have a severe hearing impairment (requiring a hearing aid on a daily basis) and that they should not have a visual impairment uncorrected by a visual aid compatible with a VR headset.

2.3.2 Experimental setup

2.3.2.1 Procedure

The experiment took place in a hemi-anechoic room (the Spatial Audio Lab at the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), Montréal, Québec). Participants wore a VR headset and completed a series of speaking tasks in immersive virtual environments, where both visuals and acoustics were manipulated and the sequence of the audio-visual environments was randomized. Participants were given a brief familiarization period before data collection began. During this time, the experimenter ensured that the VR headset was properly fitted and that participants could see the images clearly and felt comfortable using the device. This step was particularly important given that some participants were experiencing VR for the first time. No major issues were reported during this phase, and all participants proceeded with the task once they confirmed they were ready.

During the speaking tasks, participants would face towards the avatar of a person called Mario at a same distance, and they were instructed to address him directly to explain or describe something for each experimental condition; this is done in the context of prompting the participant to talk to a newcomer to Montréal, who would like to know the important how-tos in a consecutive, story-like manner. For example, how to prepare a poutine (a French Canadian recipe), how to get to the airport, or how to wash a car. The participants were instructed to speak for at least

one minute for each task, and as a control for the experiment, Mario played a role of a passive listener and did not provide any auditory or visual responses to the participants' explanations. The instructions, 10 to 15 s each, were pre-recorded and played in the room; during the time the instructions were playing, the VR headset displayed a neutral, single-color space. After each instruction, participants were immersed in the following VR scenario. Both the avatar and the participants were standing, and participants were instructed to limit physical movements and especially to remain centered to the microphone (visible to the participants) at a constant, comfortable distance. After all conditions, there was an interview on the participants' experience during the experiment. The qualitative interview covered the participants' overall impressions of the experiment, specific feelings about comfort, pleasantness, and difficulty in different audiovisual environments, and their self-perception of voice production during the experiment.

2.3.2.2 Immersive virtual environments

There were three visual environments and three acoustic environments; this resulted in nine audio-visual conditions from the three-by-three combinations. The three visual environments realized in VR were a gymnasium, a classroom, and a hemi-anechoic room; they were rendered from existing rooms using a 360-degree camera, with the hemi-anechoic room being the room where the experiment took place. Figure 2.1 shows an example of the VR rendering, and Table 2.1 shows the dimensions of the rooms and their corresponding reverberation level. These three rooms were visually distinct and had different room acoustics.

Table 2.1 The dimension and the reverberation level of the three rooms

	Width (ft)	Length (ft)	Height (ft)	Volume (cu ft)	Reverberation Level
Hemi-Anechoic Room	18	21	12	4,536	Low
Classroom	24	30	10	7,200	Middle
Gymnasium	114	116	31	409,944	High

The acoustic environments were created by a diffuse room impression for the high and medium reverberation conditions, with the impulse response recorded in each room. A shotgun microphone (Neumann Model KMR81) was positioned in the middle of the width and about one

Figure 2.1 Example VR rendering of the classroom

third in from the length of the hemi-anechoic room, where the experiment took place. The mono signal from this microphone was then received in an audio interface (Apogee Model Duet) and connected to a laptop (Apple Model MacBook Pro A1286, Cupertino, California, United States). In a digital audio workstation (DAW Reaper) this signal was then convolved in real time with a mono impulse response previously taken in the gymnasium or the classroom. Then, a second, phase-reversed copy of the convolved signal was created, and the resulting stereo feed was sent to two loudspeakers (Genelec 8020A, Iisalmi, Finland) placed in the opposite corner of the room, facing away from the shotgun microphone. On the backside of the two loudspeakers and in between the loudspeakers and the shotgun microphone an absorptive wedge of the anechoic room was placed so as to attenuate and minimize the direct sound that would create a feedback loop. Effectively, participants heard their voices convolved with the corresponding reverberation in different acoustic environments played back from the loudspeakers in the experimentation room. For the low reverberation condition, it was the room acoustics of the experimentation room itself. Throughout all experimental conditions, participants experienced both visual and auditory environmental information simultaneously, as they would in natural speaking environments. Our 3×3 factorial design manipulated room type independently for each sensory modality (visual

room type and auditory room type), allowing us to estimate the contribution of each factor and the interactions between them.

2.3.3 Data preprocessing

The SPLs from the shotgun microphone were measured with Praat (Boersma & Weenink, 2001) at 20 ms intervals. Further data processing was done in R (Team, 2022) with its stats package (for LOESS) and the dplyr package (Wickham, François, Henry & Müller, 2022). We removed the silences and non-sonorant measurements by examining the SPL distribution; the cut-off point was 35 dB higher than the noise floor. SPLs in speech vary largely when measured at this sampling rate, and this variation adds noise to the true trend of the overall SPL trajectory. Thus, we then calculated a moving average SPL curve for each recording with local polynomial regression using the locally estimated scatterplot smoothing (LOESS) method, accounting for its nonlinearity. Only the first 30 seconds of recording was considered. The beginning of the utterance in each recording was manually extracted, serving as the reference start time. In total, we had 190,882 data points, corresponding to on average 1055 data points per recording. Figure 2.2 shows an example of this data preprocessing procedure. Lastly, similar to Bottalico *et al.* (2016), we performed by-participant centralization for SPLs with the mean calculated from all the data points for each participant to maintain the difference between different conditions.

2.3.4 Data analysis

We used GAMM to model the SPL trajectory over time in each audiovisual condition. GAMM is a powerful statistical technique for modeling data that may have a non-linear pattern. The model's "additive" components build a flexible curve by combining multiple simpler, curved shapes. The complexity of this curve is controlled by a parameter called the number of knot (k). A higher number of knots increases the flexibility of the fitted curve. However, an excessive number can lead to overfitting, causing the model to capture noise rather than the underlying pattern. We used the "mixed" version of the model to account for the unique characteristics of our participants, as our data was clustered by participant from repeated measures. We chose

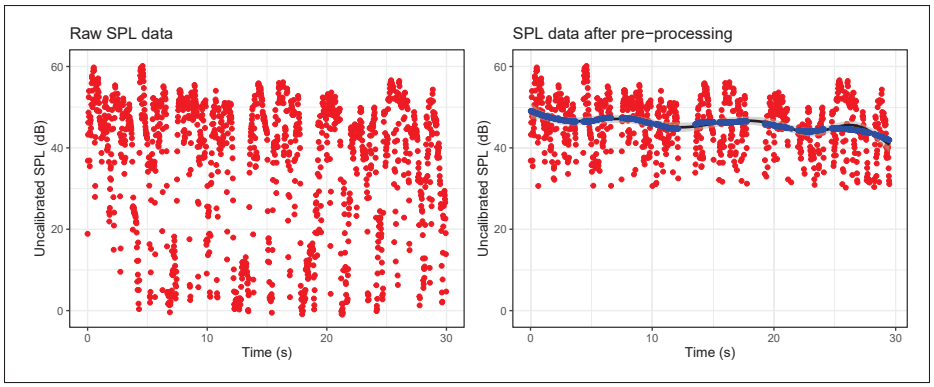


Figure 2.2 An example of the data pre-processing treatment from one recording

The left-hand sub-plot shows the raw data, and the right-hand sub-plot shows the data after pre-processing. The red points are individual data points; the blue points are the LOESS-treated data points; and the black curve is a default optimized smooth curve for this recording from the geom_smooth function in the ggplot2 package (Wickham, 2016).

GAMM for analyzing our data because the SPL trajectories are nonlinear and different conditions may have a different underlying shape, which can be modeled with the flexibility that GAMM offer. We built our models in R (Team, 2022) with the mgcv package (Wood, 2017).

The main model is built as below, where Time stands for the timestamp of the data point, Aud stands for the auditory condition, Vis stands for the visual condition, and Participant stands for the participant ID:

$$\begin{aligned} SPL \sim\ & s(Time) \\ & + s(Time, Aud, bs = \text{“}sz\text{”}, k = 5) \\ & + s(Time, Vis, bs = \text{“}sz\text{”}, k = 5) \\ & + s(Time, Aud, Vis, bs = \text{“}sz\text{”}, k = 5) \\ & + s(Time, Participant, bs = \text{“}fs\text{”}, m = 1, k = 5) \end{aligned} \tag{2.3}$$

The five smooth terms model below rewrites the model above and explains what each term captures:

$$
\begin{aligned}
SPL \sim\ & \text{Overall trend over Time} \\
& + \text{Effect of Aud over Time} \\
& + \text{Effect of Vis over Time} \\
& + \text{Interation effect of Aud and Vis over Time} \\
& + \text{Random smooth effect for each participant}
\end{aligned}
\tag{2.4}
$$

In our model, we have two categorical independent variables (Aud and Vis) that can interact with each other, and a continuous variable (Time) that can interact with the two categorical variables.

From this model structure, we extracted partial effects and interaction effects to understand how each factor influences SPL trajectories. The partial effects (referred to as main effects in ANOVA or linear mixed effects modeling terminology) represent the contribution of each factor, averaged across all levels of the other factors. In our model, this includes the overall effect of Time, which captures the general SPL trajectory; the effect of auditory condition over time, which captures how different auditory environments modulate the SPL trajectory; and the effect of visual condition over time, which captures how different visual environments modulate the SPL trajectory. The interaction effect captures how the influence of auditory and visual conditions on SPL trajectories depends on specific combinations of the two factors—that is, whether the effect of one modality varies depending on the level of the other modality.

To model these interaction effects, we used the sum-to-zero smooth interactions (bs = "sz"). This makes sure that the two categorical variables are modeled as orthogonal to each other, capturing each of their effect on SPL. The "sz" option also creates one curve per level of the categorical variable and models the difference between each level to the overall trend captured by the continuous variable (Wood, 2017). It highlights the effects of Aud and Vis on SPL over time by removing the overall partial effect of Time. The number of knot function (k = 5) was

chosen by iterative inspections, aiming to capture the overall changes but is not too sensitive to the rapid fluctuations in SPLs that are not meaningful to the current research.

To test the significance of the effects of Aud and Vis and their interaction, we also built a series of reduced models that removed these terms while keeping the random effect term constant. Specifically, from the main model shown above, we created four reduced models: 1) a model with the Aud and Vis partial effects, but without their interaction; 2) a model with only the Aud partial effect; 3) a model with only the Vis partial effect; 4) a null model containing only the overall effect of Time. We conducted model comparisons to determine if including these terms significantly improved model fit. The core principle behind this approach is to test whether a given term makes a meaningful contribution to explaining the data. By comparing a model that includes a term to a simplified version that excludes it, we can assess if it is beneficial to consider that term. The comparisons used Akaike information criterion (AIC) and chi-squared tests on the difference in the minimized smoothing parameter selection scores between models.AIC works by scoring each model based on a trade-off between goodness of fit and simplicity. The AIC score is calculated from two components: how much of the data the model fails to explain (or in a way, the model's "badness" of fit) and a penalty for model complexity, which is determined by the number of parameters used. The absolute value of the AIC is not interpretable because its value is highly dependent on the dataset and the scale of the model's likelihood function. A high or low AIC in isolation is not meaningful. Thus, AIC is used solely for comparative model selection among candidates fit to the same dataset. When comparing two models, a lower AIC indicates a better model and a p-value lower than 0.05 in the chi-squared test suggests a statistically significant difference between them.

Moreover, using the gratia package (Simpson, 2024), the ggplot2 package (Wickham, 2016) and the itsadug package (van Rij, Wieling, Baayen & van Rijn, 2022), we extracted and visualized the estimated smooth curves for SPL with the 95% confidence interval (CI) from the best-fit GAMM model to further investigate the effects of interest.

Finally, we examined how different audio-visual conditions influenced speakers' initial rate of change in speech level. From the GAMM-fitted curves (N=180 with 20 participants × 9 conditions), we quantified the initial rate of change by computing the slope between two key points: (1) the speech level onset (initial point) and (2) the first inflection point in each curve. This slope was calculated as $\Delta SPL/\Delta time$, representing the rate of dB change per second. Using the lme4 package (Bates, Mächler, Bolker & Walker, 2015) , we then fitted a series of linear mixed-effects models (LMEs). We began with a null model (intercept only) and incrementally added fixed-effect terms while keeping the participant-level random intercept. Each more complex model was compared to its predecessor using likelihood ratio tests (LRTs) to assess whether the added term significantly improved model fit. Post-hoc pairwise comparisons were conducted using the emmeans package (Lenth, 2022).

2.4 Results

2.4.1 GAMM model comparisons

Comparing the main model to a model without the interaction term between Aud and Vis, the AIC was 1761.2 lower in the main model ($p < .001$). Continuing in the reduction, we compared models without Aud or Vis to the model with only the interaction term removed to test the significance of the partial effects of Aud and Vis. Results showed that both partial effects were significant ($p < .001$); removing Aud led to 26156.22 increase in AIC, while removing Vis led to 2123.7 increase. The complete model comparison results are summarized into a table provided in the Supplementary Material. The large AIC values and changes in AIC are expected due to the high sampling frequency, which resulted in a large dataset.

In summary, the model construction provided in Section 2.3.4 was the best-fit model; the results below are from the outputs of this model.

2.4.2 Estimated smooth curves for all Aud-Vis conditions

The estimated smooth curves for the Aud-Vis conditions grouped by visual condition and auditory condition are shown in Figure 2.3 and Figure 2.4 respectively. In general, we use color to code different conditions: black for the hemi-anechoic room (Anechoic, or A), blue for the classroom (C), and cyan for the gymnasium (Gym, or G). We use line type to indicate different modalities: solid lines represent visual conditions and dashed lines represent auditory conditions. Using the same color for auditory and visual conditions from the same source environment best illustrates their connection.

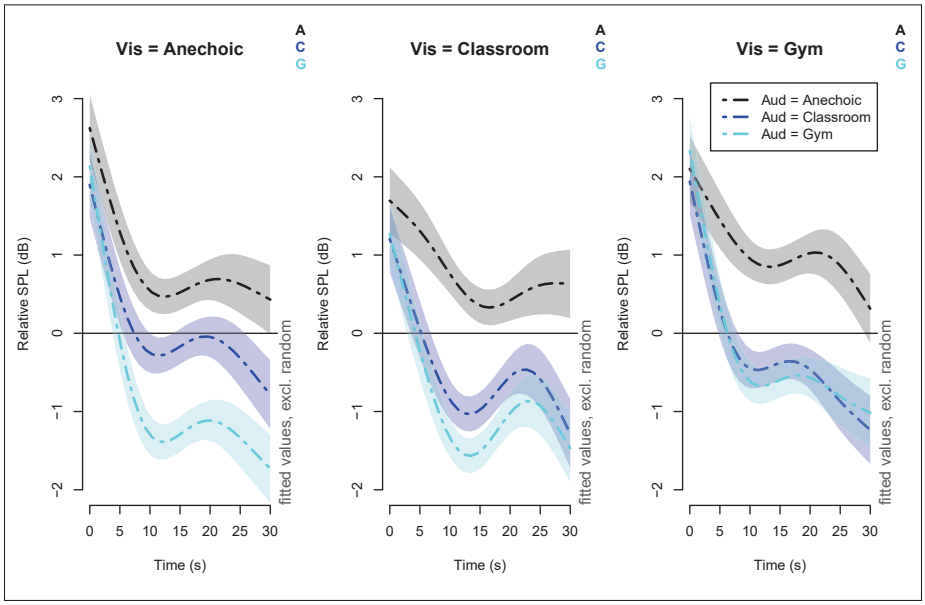


Figure 2.3 The estimated smooth curves of all the Aud-Vis conditions grouped by the visual conditions

The visual condition is listed on the top of each sub-plot, and the auditory conditions are shown in the legend. X-axis shows time in seconds, y-axis shows centered SPL in dB as described in Data processing 2.3.3. The lines show the model estimates, and the shades show the 95% CI.

Overall, the dynamic range of speech levels over time in different conditions is 2 to 4 dB. We see that overall the curves showed three phases: rapidly decreasing for around 10s, increasing or stagnating for around 10s, and then decreasing for another 10s. Overall, we see differences

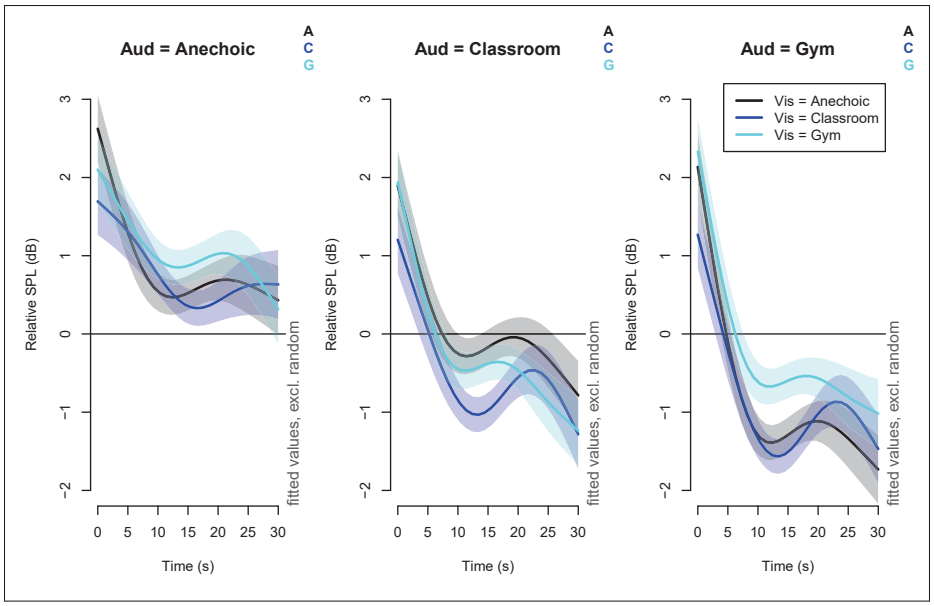


Figure 2.4 The estimated smooth curves of all the Aud-Vis conditions grouped by the auditory conditions

The auditory condition is listed on the top of each sub-plot. The color codes different visual conditions.

in the effect of Aud in different visual conditions and vice versa, although there are more overlaps between the three curves on each sub-plot in Figure 2.4 than in Figure 2.3. As shown in Figure 2.4, the most rapid change of speech level is seen when Aud is the gymnasium (G for short), followed by the classroom (C for short) and the hemi-anechoic room (A for short).

2.4.3 Partial effects

2.4.3.1 Time

The partial effect of time on speech levels is shown in Figure 2.5. Overall, regardless of the Aud and Vis effects, the speech level decreased significantly by approximately 2.5 dB within 14 seconds, and then fluctuated within 1 dB with a tendency of decreasing.

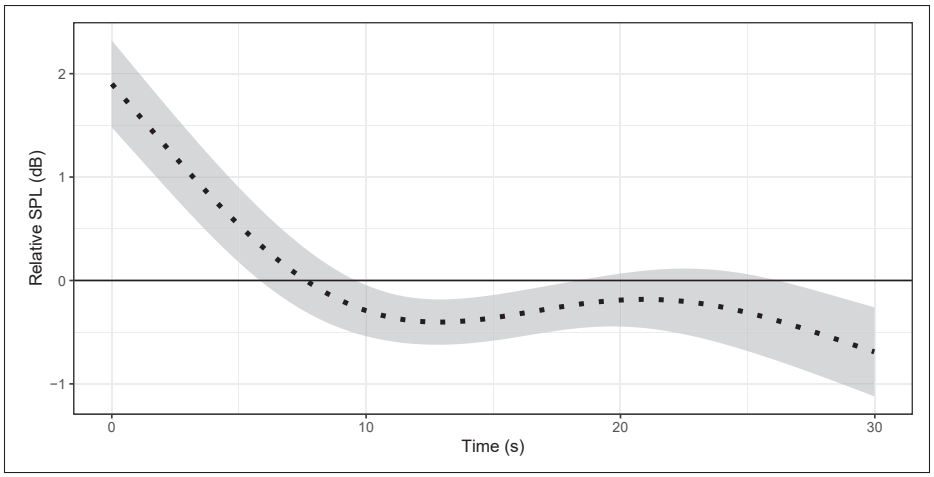


Figure 2.5 The partial effect of Time on speech levels

2.4.3.2 Auditory immersion

Figure 2.6 shows the partial effects of Aud on SPL over time. As a reminder, the estimated smooth curves are relative to the partial effect of time on SPL. Overall, the effects of Aud range from -0.8 to 1.25 dB over time. At the start, the three curves did not differ much and were all close to zero. Within the fist 10s, SPLs relatively increased in Aud condition A (Aud-A for short), decreased in Aud-G, and stayed the same in Aud-C. After 10s, we see some fluctuations in SPLs but overall it was stable.

2.4.3.3 Visual immersion

Figure 2.7 shows the partial effects of Vis on SPL over time. Overall, the effects of Vis range from -0.6 to 0.4 dB over time, and we see large variations over time. At the beginning, the SPL curves from Vis-A and Vis-G started around the same position at 0.25 dB, while the SPL curve from Vis-C started at -0.5 dB. Despite the small magnitude of this difference, these starting points were statistically significantly different. The Vis-A curve then decreased to 0 dB within 10s and fluctuated around it; the Vis-G curve increased and peaked at around 15s, and then decreased to around 0 dB; the Vis-C curve increased with large fluctuations and also ended at around 0 dB. Similarly, the middle section where the curves diverged showed statistically

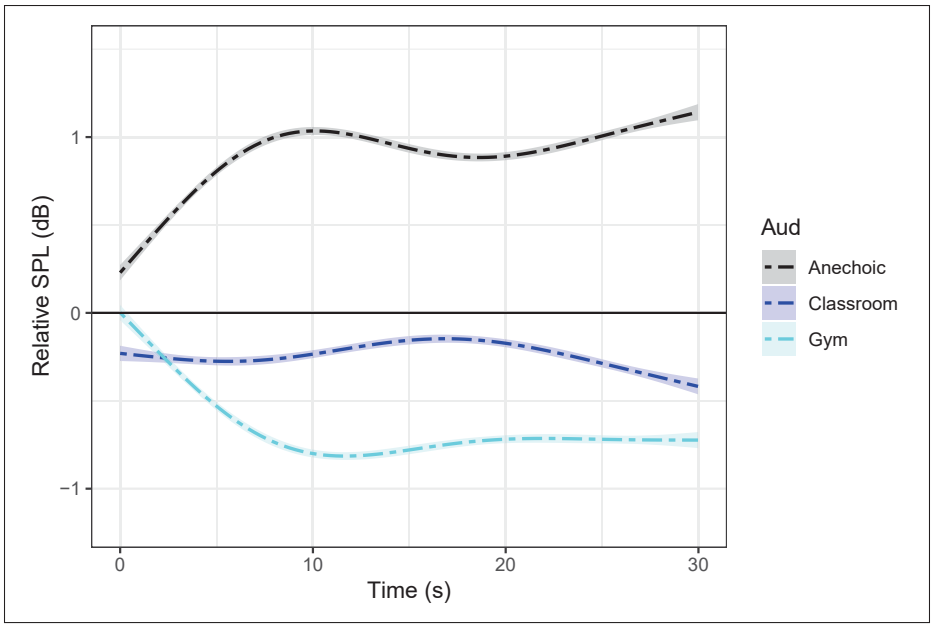


Figure 2.6 The partial effects of auditory immersion on SPL over time

The color codes the auditory conditions.

significant differences from one another. However, once the three curves converged, they were no longer statistically different.

2.4.4 Interaction effects of auditory and visual immersions

2.4.4.1 Effects of auditory conditions in different visual conditions

First, we examine the effects of auditory conditions given each visual condition. We plotted the nine curves in a way that highlights these differences in Figure 2.8. As a reminder, these interaction effect curves show the changes in SPLs in additional to the partial effects of Time and the partial interaction effect between Time and Aud or Time and Vis.

Overall, Aud conditions further influence SPLs differently in different visual conditions, even though the magnitude range is small (-0.4 to 0.4 dB). In Vis-A, the three Aud curves started around the same position, then Aud-C increased and stabilized at 0.3 dB while Aud-G continuously

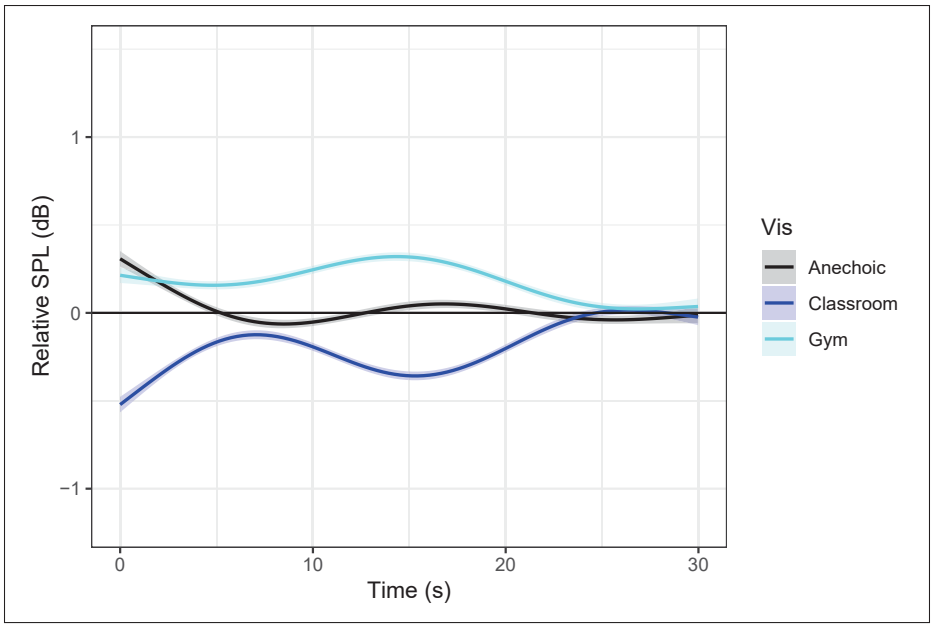


Figure 2.7 The partial effects of visual immersion on SPL over time

The color codes the visual conditions.

decreased to -0.3 dB and Aud-A decreased and bounced back to around 0 dB. In Vis-C, the three Aud curves were intertwined with each other, not showing clear differences. In Vis-G, we see that Aud-G curve is higher than the other two overall. Aud-C curve decreased from 0 dB to -0.2 dB with fluctuations, and Aud-A curve stagnated in between Aud-G and Aud-C around 0 dB with fluctuations.

Looking across the three sub-plots from left to right, we see that the effect of Aud-G went from overall negative to overall positive, while the effect of Aud-C was the reverse.

Figure 9 further accentuates the differences between Aud-C and Aud-G in different visual conditions. We focus on this contrast as it revealed the most consistent and interpretable pattern across conditions. Here, unlike the differences we saw in Figure 2.8, which were solely the difference in the interaction effects, the difference curves in Figure 2.9 were calculated from the overall trend of Aud-C and Aud-G in the three visual conditions. We can see that these three curves started around the same position at 0 dB and then diverged, with Vis-A showing the

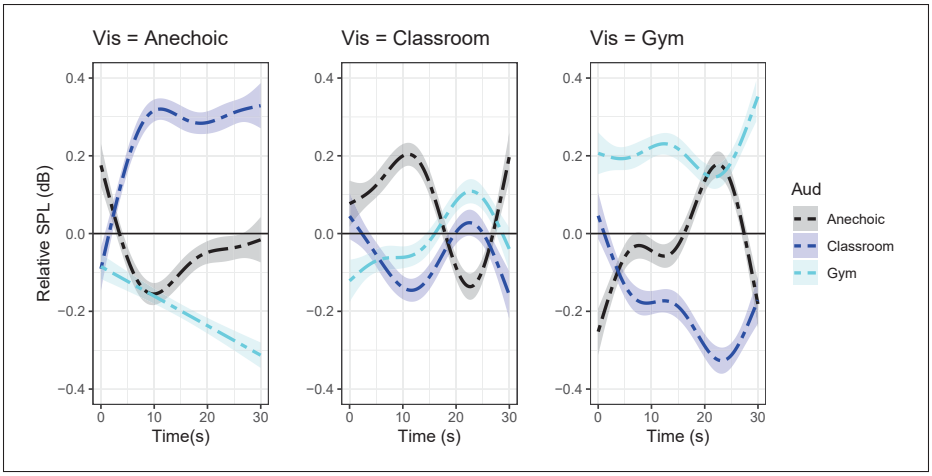


Figure 2.8 The effects of auditory conditions in different visual conditions

Each sub-plot is one Vis condition, and each color is one Aud condition.

largest difference, followed by Vis-C and Vis-G. In Vis-G, parts of the curve were below 0 dB, showing that Aud-G was higher than Aud-C at the beginning (before 6s) and the end (after 23s). Overall, we can identify two phases in these curves: before around 10 to 15 seconds, the curves went up, and after, the curves either plateaued (as in Vis-A) or went back down.

2.4.4.2 Effects of visual conditions in different auditory conditions

Next, we examine the effects of visual conditions given each auditory condition, as shown in Figure 2.10. Again, we see overall differences in how the visual conditions influence SPLs in different auditory conditions. When the auditory condition was A, the three Vis curves were intertwined. When the auditory condition was C, all three curves started around 0 dB and then diverged, with Vis-A increasing, Vis-G decreasing, and Vis-C fluctuating in the middle. When the auditory condition was G, Vis-G had the highest SPLs, followed by Vis-C and Vis-A.

Similar to Aud-C and Aud-G in different visual conditions, we see that Vis-A and Vis-G also switched in their relative positions across the three auditory conditions —a pattern that stood out consistently. To further exanimate this difference between Vis-A and Vis-G, we plotted

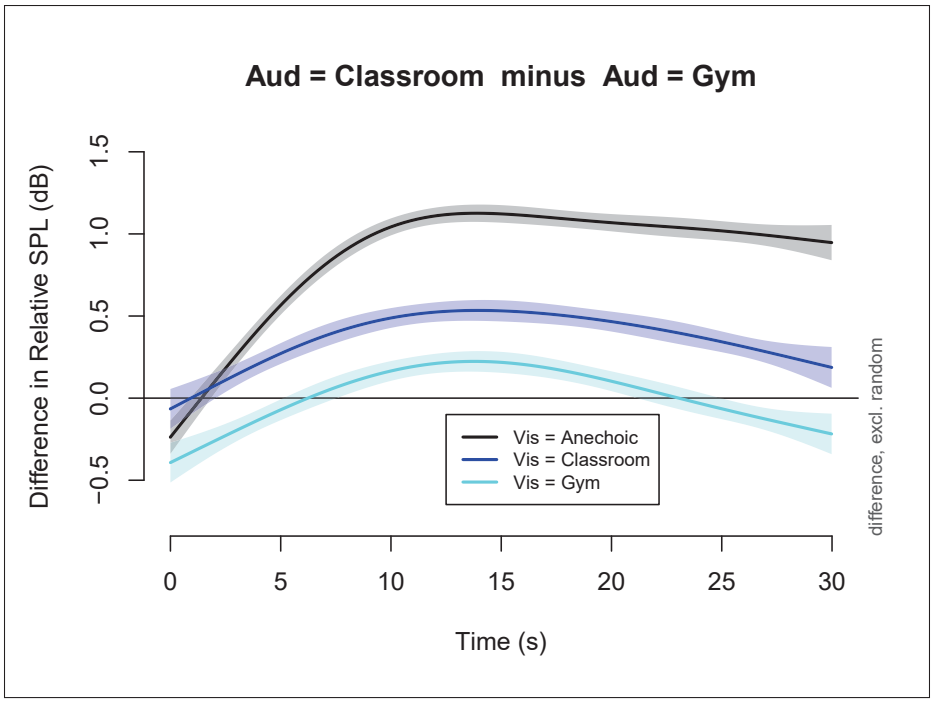


Figure 2.9 The difference curves between Aud-C and Aud-G in the three visual conditions

The structure of the figure follow the other ones. For example, the black curve is the difference between the estimated Aud-C Vis-A curve and Aud-G Vis-A curve.

their difference curves for the three Aud conditions in Figure 2.11. The Aud-C and Aud-G curves started around the same position (0.1 dB) and then diverged, with Aud-C decreasing and Aud-G increasing. The Aud-A curve started around -0.5 dB, increased in the first 10s, plateaued until 22s, and then decreased again. Overall, the difference between Vis-G and Vis-A was the largest when the auditory condition was G, with Vis-G having higher SPLs than Vis-A. When the auditory condition was A, Vis-G at first had lower SPLs than Vis-A, but this difference kept decreasing until it was reversed during the first 10s; after 22s, the trend reversed again. When the auditory condition was C, Vis-G always had lower SPLs than Vis-A.

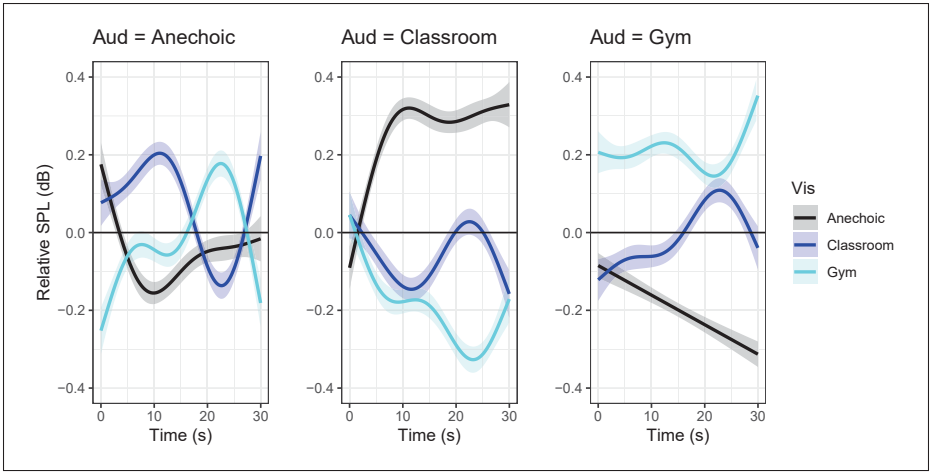


Figure 2.10 The effects of visual conditions in different auditory conditions

Each sub-plot is one Aud condition, and each color is one Vis condition.

2.4.5 Initial rate of change in speech level

From the sequential model construction, the addition of AudCondition significantly improved model fit compared to the null model ($p < .001$), whereas VisCondition showed only marginal improvement ($p = .061$). The additive combination of both factors provided significantly better fit than either single-factor model (vs. AudCondition-only: $p < .05$; vs. VisCondition-only: $p < .001$). However, including an interaction term did not further enhance the model fit ($p = .460$), supporting an additive rather than interactive relationship between conditions. The final model therefore included both AudCondition and VisCondition as independent predictors with by-participant random intercept, explaining significantly more variance than simpler alternatives while maintaining parsimony.

The optimal LME revealed significant effects of audio-visual conditions on the speech level's rate of change. Figure 2.12 provides a visual summary of the fixed effects of the best-fit model with the effects package (Fox & Weisberg, 2019). For the auditory conditions, the estimated initial speech level rate of change was -0.049 dB/s for Aud-A, -0.097 dB/s for Aud-C, and -0.152 dB/s for Aud-G. For the visual conditions, the estimated initial speech level rate of change was

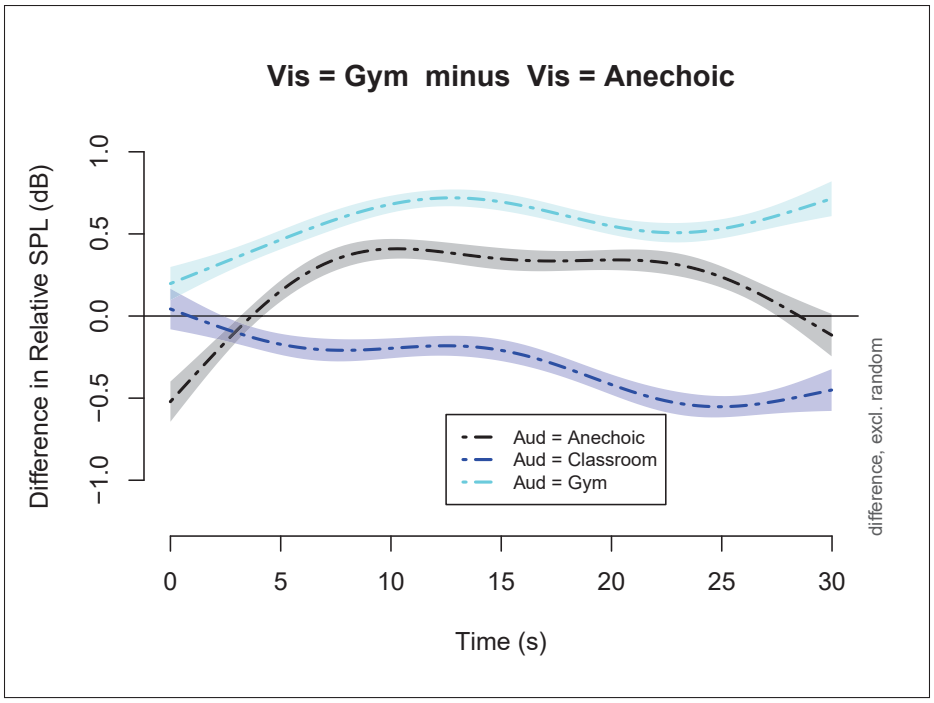


Figure 2.11 The difference curves between Vis-G and Vis-A in the three auditory conditions

The structure of the figure follows the other ones. For example, the black curve is the difference between the estimated Aud-A Vis-G curve and Aud-A Vis-A curve.

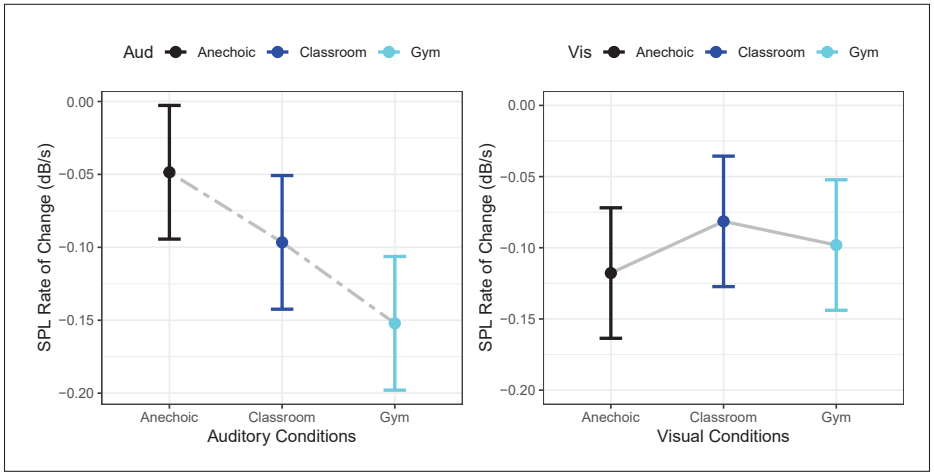


Figure 2.12 The effects of auditory and visual conditions on the speech level rate of change.

-0.118 dB/s for Vis-A, -0.081 dB/s for Vis-C, and -0.098 dB/s for Vis-G. Pairwise comparisons with Tukey adjustment demonstrated that Aud-G produced significantly higher rates than both Aud-A (mean difference = 0.104, 95% CI [0.073, 0.134], $p < .001$) and Aud-C (mean difference = 0.056, 95% CI [0.025, 0.086], $p < .001$), while Aud-C also showed significantly higher rates than Aud-A (mean difference = 0.048, 95% CI [0.017, 0.079], $p = .001$). For visual conditions, only Vis-C showed significantly higher rates than Vis-A (mean difference = 0.036, 95% CI [0.006, 0.067, $p = .016$), with no other significant pairwise differences observed.

2.5 Discussion

2.5.1 Comparing the overall auditory and visual effects

The auditory partial effect showed that participants spoke the loudest in Aud-A, followed by Aud-C and Aud-G. This sequence reversely correlates with the levels of reverberation in the three Aud rooms. However, there could be different explanations as to why participants changed their speech levels in this way. One possibility is that participants' voices were more supported in the more reverberant room, and thus they did not have to speak as loud. Although, it could also be that participants did not like hearing the echoic reverberation of their own voice, and thus they decreased their speech levels.

The fact that the difference between the three curves increased over time from a similar starting point shows two things: 1) Participants started speaking at a certain level (potentially a level of habit or comfort) and then adjusted to match the auditory conditions; 2) The effect of the auditory condition is gradual—it took around 10s for the speech levels to become stable. The duration of the adjustment period is surprising. Auditory feedback control on speech production often shows effect within 2s (e.g., Villegas *et al.* 2021), while auditory perturbation studies with pitch or formants show even faster responses within hundreds of milliseconds (e.g., Purcell & Munhall 2006; Xu, Larson, Bauer & Hain 2004). To our knowledge, though, no existing studies have directly compared how people react to noise or other auditory perturbations versus reverberation. One likely explanation is that people adapt faster to noise because it has a more obvious effect

on how clearly they can hear their own voice; auditory perturbations of F0 and formants involve even more direct modification of one's speech production, which may explain the fastest response times. Reverberation is less noticeable in comparison, so this adjustment likely takes longer to develop. Another factor may be that the other types of perturbations are constant compared to reverberation which only shows its effect when a sound is made (such as speaking) in the environment. Since natural speech includes pauses and silences, this makes reverberation even harder to adjust to. Future studies could directly compare these conditions in the same experiment to determine exactly how people react to them differently and identify the underlying causes.

On the other hand, the visual partial effect was distinct from the auditory one. First, unlike in the auditory partial effect, all three curves converged at the end, meaning that the visual differences played less of a role over time, and there was a relatively larger difference between the three curves at the beginning. This shows that the visual effect is more instantaneous, which makes sense since the visual information is already present, while knowing the acoustics of a room requires sound production.

What is surprising is the relative positions of the three visual conditions. At the beginning, Vis-G and Vis-A started around the same point, even though they were the most visually different. In the middle section, participants spoke the loudest in Vis-G, followed by Vis-A and Vis-C. This sequence cannot be explained simply—it does not strictly correlate to the size of the rooms nor the (expected) room acoustics. In fact, G has the highest speech level in visual conditions yet the lowest in auditory conditions. One possible explanation is that it is related to how people are used to speaking in these rooms. People are used to speaking loudly in a gym and being more quiet in a classroom, regardless of the size or the acoustics; the participants did not have much experience with a hemi-anechoic room aside from speaking in this room before the experiment started, and potentially, they increased their levels when seeing Vis-A due to the room acoustics association. The hypothesis that participants may have spoken according to their habitual patterns for a given room type – what one of our reviewers termed the "semantic association of environment" effect or simply "room semantics" effect – needs further testing by

comparing these specific, common speaking environments to visually ambiguous environments that vary only in size. A recent study by Nudelman & Bottalico (2025) provides relevant insights. Researchers tested participants in three distinct conference room sizes, each with two versions: sparsely occupied and densely occupied. Since all rooms were conference rooms and thus have similar room semantics, the room occupancy could serve as a factor for semantic differentiation. While the results showed no significant SPL differences between occupancy conditions but only room size (consistent with Daşdöğen *et al.* 2023), the densely occupied rooms did lead to higher vocal fatigue and discomfort ratings. However, this study also had the same limitation as Daşdöğen *et al.* (2023): communication distance varied with room size. Future studies examining room semantics association effects should also control for communication distance to isolate the specific influence of room semantics on vocal behavior.

While both Aud and Vis had statistically significant effects on speech levels, the large difference in AIC when testing the partial effects of Aud and Vis showed that Aud has a stronger effect than Vis. This has also been shown in the magnitude of difference in speech levels in the estimated curves comparing Figure 2.6 and Figure 2.7. This is within expectation as the room acoustics directly affects one's auditory feedback control and the visual effect has been shown to be more short-lived. However, in previous studies (Daşdöğen *et al.*, 2023; Nudelman & Bottalico, 2025) where communication distance was not controlled, the visual effect of changing room size had a larger magnitude that was more comparable to the auditory reverberation effect. Another potential explanation for this limited visual influence could be related to the cognitive demands of the spontaneous speech task. As noted by a reviewer, generating spontaneous responses requires real-time content formulation, which may have diverted participants' attention from the visual input and reduced their sensitivity to the multisensory conditions. Future research could explore this by comparing spontaneous speech tasks with simpler speech tasks that minimize cognitive load, allowing for a clearer assessment of visual effects on vocal behavior.

When examining the initial rate of change in speech level, both auditory and visual effects were statistically significant. Consistent with the GAMM results, the auditory effect was stronger than the visual effect. In the three auditory conditions, the rate of change increased with the level of

reverberation, suggesting that the adjustment in speech level was not made at a constant rate irrespective of the acoustic environment. Instead, participants appeared to adjust their speech level differently depending on the acoustic context, changing more quickly in environments that required a larger adjustment. In the visual conditions, the rate of change also varied, with the classroom condition showing the slowest increase in speech level. This further supports the influence of visual context on speech level production. Future research should systematically explore the mechanisms underlying these effects. Given the relatively smaller and more variable effects observed especially across visual conditions, future research would benefit from including more trials per visual condition to improve reliability and sensitivity.

Lastly, our factorial design allowed us to estimate how auditory room characteristics and visual room characteristics each contribute to speech level adjustments in a multisensory context. We cannot speak to how speakers would behave if one modality were entirely absent or uninformative, but our findings characterize how each type of environmental information influences speech level control when both are available, as in naturalistic speaking situations.

2.5.2 Trends in the Aud-Vis interaction effects

In Section 2.4.1 and Section 2.4.4, we established that there are significant interaction effects between the auditory condition and the visual condition, and in Section 2.4.2, we saw that the nine curves differ from one another. In this section, we would like to further explore what these interactions mean in terms of audio-visual integration.

In Figure 2.9, we showed the speech level difference in Aud-C and Aud-G in different visual conditions, and the trend we saw can be explained with audio-visual integration. As shown in Figure 2.6, Aud-C and Aud-G were two auditory conditions that bear more similarity—indeed, they both had added reverberation. What we see in Figure 2.9 showed that this difference varies according to visual conditions. When participants see the hemi-anechoic room, the difference between Aud-C and Aud-G is the largest; we may treat this visual condition as the reference, since this was the physical room that the participants were in. When seeing the classroom, the

visual effect of the classroom was amplified, and the speech levels decreased for Aud-C. When seeing the gym, again, the visual effect of the gym was amplified, and the speech levels increased for Aud-G; these movements are more visible in the estimated curves for all conditions plotted in Figure 2.3.

We also saw an analogous visual difference effect by Aud. In Figure 2.11, the difference between Vis-A and Vis-G in each auditory condition was plotted. Again, these were the two conditions that were more similar to each other, empirically speaking (as shown in Figure 2.7). The difference between Vis-A and Vis-G was the most unstable, suggesting difficulty in speech level control, and overall with the smallest magnitude. Overall, Vis-G had higher speech levels than Vis-A, as manifested in the visual partial effect. Their difference was amplified when Aud was G, with Vis-G increased and Vis-A decreased as shown in Figure 2.10. When hearing the classroom, their relationship was reversed, with Vis-A increasing and Vis-G decreasing; this fits with the overall auditory effect of A and G as shown in Figure 2.6, suggesting that participants potentially were adjusting their speech levels according to the expected acoustics of the room they saw. It is fitting that this is observed with Aud-C, because Aud-C has the least effect on speech levels over time, serving as the empirical reference level for the participants instead of the designed reference (i.e., Aud-A). This is echoed from the post-experiment debrief interview: Many participants reported that they did not know why "sometimes the microphone was working, sometimes it wasn't" and "sometimes there was a lot of echos"—they were associating hearing themselves through the medium reverberation condition (i.e., Aud-C) to "microphone working" as the "norm".

When examining the interaction effects between Aud and Vis, we noticed some trends and similarities in Figure 2.8 and Figure 2.10. We calculated the pairwise Euclidean distances between all nine interaction-effect curves, whose values ranged from 0.5 to 4.9. Interestingly, the closest matches emerged when the Aud-Vis conditions were comparable, i.e. when the auditory and visual conditions either matched or differed by one level in terms of their effective impact on the signal. This finding may also support Aud-Vis integration in speech level control. Detailed analysis is presented in the supplementary material.

2.6 Conclusion

In this research, we investigated the effects of auditory and visual immersions on speech levels. Our results were in line with previous works that showed that people decrease their speech levels with the increase of room reverberation. Our results also showed significant effects of the visual room on speech level; however, the levels did not strictly follow either the expected room acoustics or the size of the room, unlike what has been suggested in the literature. Overall, auditory information played a stronger role in speech level than visual information. However, the visual information had an earlier effect and then diminished, while the effect of auditory information increased and stagnated throughout. Many pieces of evidence show that people not only use both streams of information but also adapt to the combination of them systematically. In summary, this study contributes to both the practical understanding of voice production in immersive environments and the theoretical framework of multisensory integration in speech control. These insights are valuable for applications in VR and communication technology, where auditory and visual immersion can be tailored to enhance vocal performance and naturalistic communication.

2.7 Author Declarations

The authors have no conflicts of interest to disclose. The data of this study are available upon request. This research project received ethics approval from the Ethics Committee for Research in Education and Psychology of the University of Montreal (CEREP-19-042-D; 180719; on 18 July 2019).

2.8 Supplementary Material

See supplementary material for the complete GAMM model comparison results, GAMM model output summary, the empirical speech level curves in every Aud-Vis condition, and the interaction curves similarity.

2.9 Acknowledgments

This research was supported by Natural Sciences and Engineering Research Council of Canada (NSERC), the NSERC CREATE OPSIDIAN Program, Fonds de recherche du Québec's Programme intersectoriel Audace (DOI: https://doi.org/10.69777/263513), CIRMMT, and the École de technologie supérieure's Marcelle Gauvreau Engineering Research Chair in Multimodal Health Monitoring and Early Disease Detection with Hearables.

CHAPTER 3

HEARING-INTEGRATED BILINGUAL SPEECH CORPUS: A FRENCH-ENGLISH CORPUS INCLUDING HEARABLES FOR STUDYING SPEECH PRODUCTION UNDER CHALLENGING LISTENING CONDITIONS

Xinyi Zhang[1,2] , Ingrid Verduyckt[3] , Rachel Bouserhal[1,2]

[1] Department of Electrical Engineering, École de technologie supérieure, Montréal, Québec H3C 1K3 Canada

[2] Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Québec H3A 1E3 Canada

[3] École d'orthophonie et d'audiologie, Université de Montréal, Montréal, Québec H3N 1X7 Canada

Paper submitted for publication in "The Journal of the Acoustical Society of America" on August 27, 2025.

3.1 Abstract

Communication challenges are exacerbated by noise, particularly for individuals with hearing impairment who may also have ear occlusion from hearing protection devices in occupational settings. These combined effects on speech production are understudied, despite auditory feedback being crucial for speech motor control and thus effective communication. This paper introduces HIBiSCus, a comprehensive database for examining speech production across varying levels of noise, ear occlusion, and hearing impairment. We recruited 49 participants (19 with at least one frequency at ≥ 20 dBHL) who completed sentence reading, sustained vowel production, and picture description tasks. We demonstrate the database's utility by analyzing speech level responses in the sentence-reading task. Using linear mixed-effects modeling, we investigated all three factors categorically and additionally modeled hearing impairment as a continuous variable using pure tone average (PTA). Key results showed that ear occlusion led to increased speech level, but the relationship was not unidirectional. Additionally, preliminary findings revealed that participants with PTA >15 dBHL spoke louder overall, and a categorical shift occurs around PTA $= 15$ dBHL, where individuals with greater hearing impairment became

less reactive to noise under high occlusion conditions. The considerable individual variability challenges traditional group-level interpretations and highlights the need for individualized modeling approaches.

3.2 Introduction

3.2.1 Hearing and speaking

During speech production, we continuously monitor our own vocal output. This largely automatic process is critical for maintaining appropriate articulation and ensuring effective communication (Levelt, Roelofs & Meyer, 1999). Many theoretical models have been proposed to formalize this monitoring mechanism (e.g., Postma 2000; Parrell *et al.* 2019); however, regardless of the specific theoretical framework, empirical evidence consistently demonstrates that auditory feedback (i.e., hearing one's own voice) plays a fundamental role in speech monitoring by providing the real-time auditory information necessary for vocal adjustments.

We hear ourselves through two pathways: AC and BC. AC transmits sound waves through the air medium, and this is also how we primarily hear external sounds, including noise in the environment and other people's voices. It effectively transmits frequencies across the audible spectrum. In contrast, BC transmits vibrations directly from the vocal tract through the skull and tissues to the inner ear. Importantly, bone and tissue act as a low-pass filter, attenuating higher frequencies and transmitting frequency components below 2 kHz (Bouserhal *et al.*, 2017b). This filtering characteristic makes the spectral content of self-perceived speech, which is typically from both pathways, different from a purely AC speech. This difference becomes apparent when people find their recorded voice unfamiliar, since recordings capture only the AC component and without the BC component through which one also hear themselves.

While the recorded voice phenomenon shows the perceptual importance of the dual-pathway auditory feedback, it does not involve real-time speech production adjustment. The speech monitoring system becomes particularly critical when auditory feedback is disrupted during

actual speech production. Although various conditions can disrupt normal auditory feedback, our study focuses on three common and interconnected factors: presence of noise, ear occlusion, and postlingual hearing impairment (which occurs after language development).

3.2.2 Noise

Background noise commonly disrupts auditory feedback during speech. In noisy environments, the AC pathway is compromised as noise masks the speaker's own voice. This masking triggers the well-documented Lombard effect, where speakers automatically increase vocal intensity to maintain audibility (Lombard, 1911; Lane & Tranel, 1971). While intensity is the most prominent modification, speakers also adjust other acoustic features including F0, formant frequency, and more (Brumm & Zollinger, 2011). Extensive research has investigated the factors that modulate the Lombard effect. A primary factor is the noise level. The Lombard effect has been observed starting from noise levels as low as 43.3 dBA (Bottalico *et al.*, 2017), though the specific rate of increase may vary across studies and conditions. Studies typically reporting intensity increases of 0.2 ~ 1 dB per 1 dB of increase in noise (Giguère *et al.*, 2006; Pearsons, Bennett & Fidell, 1977; Nijs, Saher & Den Ouden, 2008). The frequency content of noise also matters: according to Bottalico & Murgia (2023), mid-frequency noise (0.5 ~ 4 kHz) produces stronger Lombard effect than low or high frequency noise. The Lombard effect is also stronger in more communicative settings, such as interactive conversation rather than reading, highlighting its function in maintaining speech intelligibility for listeners (Villegas *et al.*, 2021; Garnier, Ménard & Alexandre, 2018). However, despite the group-level research showing consistent effects, individual differences in noise responses remain understudied.

3.2.3 Ear occlusion

Ear occlusion represents a less obvious influence on speech production compared to noise exposure, but understanding its effects both in quiet and in noise (See Section 3.2.5) has become increasingly important. This is due to the growing prevalence of ear-occluding devices like earbuds and in-ear wearable devices (hearables), HPDs, hearing aids, and communication

devices, which can be used both in quiet and in noise (Feder *et al.*, 2017; Feder, Marro, McNamee & Michaud, 2019; Seol & Moon, 2022). Ear occlusion creates two simultaneous changes in self-auditory feedback: AC speech is attenuated due to the physical blockage of the ear canal, while BC speech is amplified within the occluded ear canal. This results in speakers hearing amplified low-frequency components and attenuated high-frequency components, experiencing a "boomy" or "hollow" voice of themselves (e.g., Carle, Laugesen & Nielsen 2002). The occlusion effect can be quantified both *subjectively* through perceptual ratings on the experience or psycho-acoustic measurements and *objectively* through acoustic measurements (Hansen, 1997). The objective method measures the difference in SPL in the ear canal when the ear occlusion is present or absent, and it is recommended to use a participant's own speech utterances as the sound source in this method (Saint-Gaudens, Nélisse, Sgard & Doutres, 2022). Carillo, Doutres & Sgard (2021) have demonstrated that shallow-fitting earplugs produce greater objective occlusion effects compared to deep-fitting devices, as the larger residual ear canal volume creates more available air space which leads to more low-frequency amplification, and vice versa; in addition, silicone earplugs produce greater objective occlusion effect than foam earplugs. While it is established that ear occlusion changes auditory feedback, how this affects speech production remains understudied, particularly in quiet conditions. Existing research presents conflicting findings regarding vocal intensity changes under occlusion in quiet, with some studies reporting increased intensity (Bouserhal *et al.*, 2019; Meinke *et al.*, 2022), others documenting decreased intensity (Navarro, 1996; Tufts & Frank, 2003), and still others finding no significant changes (Tufts & Frank, 2003; Mitsuya & Purcell, 2016; Vaziri, Giguère & Dajani, 2022). These conflicting findings may reflect an important but understudied aspect of occlusion effects: the inevitable competing changes from attenuated AC and amplified BC pathways, as also noted by Vaziri *et al.* (2022). However, since previous studies report only group averages and used different ear occluding devices, we cannot determine whether the inconsistencies reflect individual variability in occlusion responses or potentially methodological differences in how various devices affect AC versus BC pathways, or other factors entirely.

3.2.4 Postlingual hearing impairment

Hearing impairment represents a chronic alteration to self-auditory feedback that can affect both AC and BC pathways across different frequency ranges. Unlike temporary perturbations from noise or occlusion, hearing impairment creates permanent changes in how speakers perceive their own speech. Lee (2012) has shown that without the presence of background noise, during a sustained vowel task that moderately-to-severely hearing-impaired speakers produce higher vocal intensity compared to normal hearing speakers, but no statistically significant difference was found in Sørensen *et al.* (2024) where participants with mild-to-moderate hearing impairment spoke in a conversation task. Existing studies have also examined and showed changes in F0 and voice quality, as reviewed in Coelho *et al.* (2015) and Di Stadio *et al.* (2025). Although, in general, research on the effect of hearing impairment on speech production—scarce to begin with compared to speech perception—has focused on the severe population, particularly those with cochlear implants, despite the American Speech-Language-Hearing Association defining hearing thresholds above 15 dBHL as hearing impairment (Clark, 1981). In addition, individual variability receives little attention, and participants are typically categorically grouped (with or without hearing impairment).

3.2.5 Connections among the three factors

These three sources of self-auditory feedback alteration not only occur commonly but also frequently co-occur, creating complex interactions that complicate the understanding of their individual and combined effects. Noise and ear occlusion are two factors that are particularly connected, as ear occlusion often serves as a protective response to hazardous noise exposure. Consequently, research has predominantly examined the combined effects of noise and ear occlusion on speech production rather than investigating the effect of ear occlusion in isolation. For example, research of speakers wearing hearing protection in noise show smaller vocal intensity increases compared to unprotected speakers, but this finding is confounded by simultaneous noise attenuation, making it impossible to isolate the effect of occlusion (e.g., Kryter 1946; Hormann, Lazarus-Mainka, Schubeius & Lazarus 1984; Tufts & Frank 2003). Another research has used

an experimental design that has only occluded-in-noise, occluded-in-quiet, and open-ear-in-quiet conditions (Bouserhal *et al.*, 2019); omitting the open-ear-in-noise condition makes it unable to directly address if ear occlusion indeed modifies one's Lombard response.

The interaction between hearing impairment and ear occlusion has received some research attention, primarily due to its relevance for hearing aid users. However, existing research has mainly focused on acoustical measurements such as the attenuation of the AC path (insertion loss) or the objective occlusion effect and perceptual outcomes such as the ratings on comfort rather than examining the effect on speech production (e.g., Winkler, Latzel & Holube 2016; Denk, Hieke, Roberz & Husstedt 2023). Given that speech production relies on auditory feedback for monitoring and control, understanding how the combination of hearing impairment and occlusion affects speakers' speech production represents a critical knowledge gap and a critical element of effective communication.

The relationship between noise exposure and hearing impairment in speech production has historical significance, as the Lombard effect was originally used as an indicator for hearing impairment detection, based on the premise that individuals with hearing impairment would show reduced involuntary responses to noise they cannot adequately perceive (Brumm & Zollinger, 2011). Contemporary research examining Lombard effects in hearing-impaired populations shows mixed and conflicting findings. One study reported that both normal-hearing and hearing-impaired speakers increased vocal intensity in noise and found no significant differences between groups in pairwise comparisons despite a significant interaction effect; the model estimates showed smaller intensity increases with increasing noise levels for the hearing impairment group, though these differences were not statistically significant (Sørensen *et al.*, 2024). On the other hand, Di Stadio *et al.* (2025) concluded that people with hearing impairment have exaggerated response to noise and that could serve as an early detection method for hearing impairment, opposite to the original observation from Lombard. Given that noise-induced hearing impairment represents one of the most common forms of hearing impairment, and affected individuals frequently communicate in challenging acoustic environments, it is important to understand this speech adaptation patterns.

3.2.6 Current study

These gaps and conflicting results show the need for a more systematic and comprehensive study that examines how these listening factors affect speech production both alone and together. Thus, we built a speech database where we manipulate noise and occlusion level while including participants across a wide range of hearing thresholds. We call it HIBiSCus. Beyond addressing the identified literature gaps, the database is designed to serve more research purposes. For example, it provides speech data in both English and French with speakers' dialectal and language nativeness information, and it has three types of speech tasks. It also includes parallel recordings from both IEM and outer-ear microphone (OEM) and a reference microphone, contributing to the limited pool of such databases (only two others exist: Bouserhal *et al.* 2019; Hauret *et al.* 2025) and advancing research in in-ear wearable technologies.

In addition to presenting HIBiSCus, we explore and demonstrate its utility through a series of analyses examining how the listening conditions affect speech level in read sentences, recorded with standard microphone placement (positioned in front of the mouth). Our primary objectives are to: (1) characterize the individual and interactive effects of noise exposure, ear occlusion, and hearing thresholds on speech level, and (2) determine whether individual variation in hearing thresholds explains differences in speakers' speech level in these altered listening conditions.

3.3 Methodology

3.3.1 Construction of HIBiSCus

3.3.1.1 Apparatus

Data collection was conducted in a double-walled audiometric booth (Eckel Noise Control Technologies, Morrisburg, Ontario, Canada; dimensions: 364 cm × 279 cm × 200 cm). The booth complies with the ANSI S12.6 (American National Standards Institute, 2008) and ISO 4869 (International Organization for Standardization, 1994) standards for acoustic uniformity,

directionality, and ambient noise levels, ensuring minimal reverberation and background noise for high-quality speech recordings.

An Interacoustics AT235 audiometer (Interacoustics, Middelfart, Denmark) was used in conjunction with Telephonics TDH-39P headphones (Telephonics, Farmingdale, New York, United States) for hearing threshold measurements. All audio equipment was connected to a Roland Octa-Capture sound card (Roland Corporation, Hamamatsu, Japan), interfaced with a Windows PC running MATLAB 2022a (MathWorks, Natick, Massachusetts, United States). Recordings were made at a 48 kHz sampling rate and 16-bit depth.

A G.R.A.S. 40HF 1" reference microphone (GRAS Sound & Vibration, Holte, Denmark), selected for its low-noise and high-sensitivity properties, was positioned 75 cm from the participant's mouth. This distance was empirically determined through preliminary trials to prevent signal clipping even during high vocal effort under noise exposure while maintaining optimal recording levels. Additionally, an intra-aural device was used, equipped with two microphones per ear: (1) an IEM, placed inside the ear canal and oriented toward the eardrum, and (2) an OEM, positioned externally to the ear canal. The earpiece incorporated FG45-32491 microphone model (Knowles Electronics, Itasca, Illinois, United States). All five microphones (one reference and four miniature microphones) recorded the vocal signals simultaneously to ensure synchronized data acquisition.

To ensure accurate SPL measurements, all microphones underwent calibration. The reference microphone was calibrated using a B&K 4231 calibrator (Hottinger Brüel & Kjær A/S, Virum, Denmark), which generated a 1 kHz pure tone at 94 dB. The recorded signal was then scaled by a calibration factor to match the reference level, establishing the microphone's sensitivity conversion from voltage to Pa. This process provided the fundamental reference for subsequent calibrations of the miniature microphones.

The miniature microphones underwent calibration to account for both sensitivity and frequency response variations. In a reverberation chamber, white noise (>85 dBA) was simultaneously recorded by each miniature microphone and a calibrated 1/4" B&K 4961 reference microphone

(Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark) positioned in proximity without physical contact. The reference microphone's signal, converted from volts to Pas using its predetermined sensitivity factor, served as the acoustic reference. Sensitivity calibration factors were derived by scaling each miniature microphone's output to match the reference pressure measurements. Additionally, the microphones' frequency responses were corrected to compensate for high-frequency attenuation caused by their placement within small slots in the earpiece assembly. This dual calibration process, addressing both sensitivity and spectral characteristics, was performed individually for all miniature microphones to ensure accurate acoustic measurements across the frequency spectrum.

The experimental setup used two devices to play sound to the participants: intra-aural loudspeakers and open-back headphones. Each intra-aural device contained a dual-driver loudspeaker assembly comprising a WBFK-30095-000 high-frequency driver (Knowles Electronics, Itasca, Illinois, United States) and a CI-22955-000 low-frequency driver (Knowles Electronics, Itasca, Illinois, United States). To play the sound without occluding the ear canals and without contaminating the reference microphone, open-back Sennheiser HD598 headphones (Sennheiser Electronic GmbH & Co. KG, Wedemark, Germany) were used, powered by a Rolls HA43 Pro amplifier (Rolls Corporation, Salt Lake City, Utah, United States) to achieve the required output levels. To ensure spectral accuracy, all the devices were calibrated using an artificial head with simulated ear canal acoustics. This approach addressed both the inherent frequency response variations of the loudspeakers and the ear canal's acoustic effects. The calibration procedure involved: (1) measuring each loudspeaker's output through the artificial head, (2) comparing these recordings to the original input signals to derive transfer functions, and (3) creating inverse filters to compensate for system-induced spectral distortions. The artificial head proved particularly valuable as it replicates human ear canal resonances and pinna effects, thereby providing realistic simulation of how sound is naturally perceived. Following filter application, verification measurements confirmed that the compensated signals matched the intended spectral characteristics when received by the artificial head.

3.3.1.2 Participants

Our study included 49 participants (19 female, 30 male) with mean ages of 35.3 and 33.3 years respectively. All participants were fluent in either French or English and met the following inclusion criteria: (1) no pre-lingual hearing impairments or voice disorders, (2) no physical hypersensitivities preventing otoscopic examination or earpiece insertion, and (3) clear ear canals confirmed via otoscopic examination. Hearing thresholds across tested frequencies ranged from -10 dBHL to 75 dBHL, with the summarized audiometric profiles shown in Figure 3.1. One participant (M45FR) has used a hearing aid frequently; we conducted an additional silent condition testing with their hearing aid.

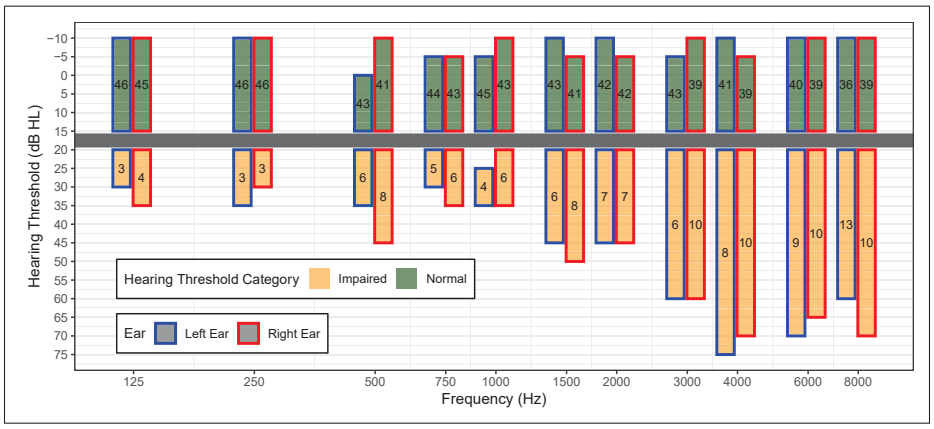


Figure 3.1 Hearing Threshold Count by Frequency and Ear

This figure shows the number of participants per ear with a hearing threshold ≤ 15 dBHL (Normal) and ≥ 20 dBHL (Impaired) at all the tested frequencies. The x-axis is the frequency in Hz and the y-axis is the hearing threshold in dBHL in a reverse scale. The number of participants is displayed at the center of each bar, while the upper and lower bounds of each bar represent the range of hearing thresholds observed at each frequency. Green bars are for ≤ 15 dBHL and yellow bars are for ≥ 20 dBHL. The blue outline of the bars stands for the left ear and red for the right ear. No data points are above 15 and below 20 dBHL due to the 5 dB step size used in the audiometric test; this gap is indicated with a grey band.

We collected detailed linguistic profiles through a questionnaire on: test language nativeness status (first language: L1; second language: L2), age of second language acquisition, self-

rated language proficiency, dialectal variation, frequency of daily use for the tested language, and multilingual status (including other known languages). Table 3.1 presents the complete breakdown of language nativeness by sex and participation language. Participants also reported any known hearing impairments, including self-assessed severity, affected frequencies, and suspected cause when known.

Table 3.1 Participant count and mean age by test language, sex, and language nativeness

Language	Sex	Nativeness	Count	Mean Age
English	Female	L1	4	43.5
		L2	5	30.2
	Male	L1	3	41.7
		L2	5	25.8
French	Female	L1	9	35.3
		L2	1	28.0
	Male	L1	20	34.5
		L2	2	28.0

3.3.1.3 Procedures and conditions

Participants were recruited through institutional mailing lists (Centre for Interdisciplinary Research in Music Media and Technology and Le Groupe de recherche en acoustique à Montréal) and word-of-mouth referrals, with explicit invitations for both normal-hearing and hearing-impaired individuals. The selection process began with an eligibility questionnaire, after which qualified candidates were invited to come to the lab for further screening. Each participant first underwent an otoscopic assessment to evaluate the tympanic membrane integrity if there were obstacles in the ear canals. This was followed by an audiometric testing measuring air-conduction thresholds at standard frequencies from 125 Hz to 8 kHz (125, 250, 500, 750, 1000, 2000, 3000, 4000, 6000, and 8000 Hz) in both ears. Only after completing these screening procedures did participants proceed to the speech recording phase of the study.

The speech production protocol for each condition consisted of three sequential tasks. Participants first read one or two sets of phonetically balanced sentences from the Hearing in Noise Test

(HINT) corpus (English: Nilsson, Soli & Sullivan 1994, French: Vaillancourt *et al.* 2005), comprising 20 sentences—the English set had 10 sentence per group, while the French set had 20 sentences per group. There were also 3 practice sentences at the beginning. These short sentences (5-7 words each) were primarily declarative statements, though we modified five sentences per condition into natural-sounding interrogative forms (e.g., “They’re going out tonight?", “Le voleur est dans la banque?") to enhance ecological validity. The same question formation was done for one of the three practice sentences. Following the sentence task, participants produced sustained versions of five vowel phonemes (/i/, /u/, /ɑ/, /æ/, /ɛ/). For French dialects lacking the /ɑ/-/æ/ distinction, participants simply repeated /ɑ/ twice. The final task involved describing three photographs from the MSCOCO dataset (Lin *et al.*, 2014) depicting real-world scenes, preceded by one practice image. This sequence (vowel production followed by image description) was repeated three times per condition. While the vowel tasks remained constant across conditions, all sentences and images were unique and appeared only once during the experiment, with their presentation order randomized across participants to mitigate order effects. To standardize speech direction and distance, participants were instructed to address an imaginary interlocutor positioned 1 meter away, with a toy head placed at this location as a visual reference point that also coincided with the screen position.

The corpus has ten acoustic conditions from combinations of three noise conditions and four ear occlusion conditions. The noise conditions included: (1) no noise - NN (quiet room), (2) low noise - LN (70 dBA), and (3) high noise - HN (85 dBA), with noise levels validated using the artificial head as previously described. For LN and HN, we varied the noise spectrum across trials: grey noise (spectrum matching typical hearing sensitivity), pink noise (equal energy per octave), and white noise (equal energy across frequencies), presented in randomized order during the sustained vowels and picture description tasks. The sentence reading task only used grey noise. The ear occlusion conditions included: (1) open ear - OE (no device, reference condition), (2) simulated open ear - SE (open-back headphones), (3) low occlusion - LO (intra-aural device with foam plugs (standard Compy^{TM} tips (Hearing Components Inc., Oakdale, Minnesota, United States)), which provide a deeper fit and thus less room for resonance), and (4) high

occlusion - HO (intra-aural device with silicone flange tips, which provide a shallower fit and thus more room for resonance). The experimental condition sequence was organized by occlusion condition, with the open-ear condition always presented first and only tested in NN condition. The remaining three occlusion conditions were randomized but followed a fixed noise condition progression: NN $\rightarrow$ LN $\rightarrow$ HN.

For intra-aural device conditions, we did device fit checks to ensure proper earpiece placement for acoustic seal by comparing sound measurements between OEM and IEM during >80 dBA white noise presentation in the room (Saint-Gaudens *et al.*, 2022; Voix & Laville, 2009). The fitting was considered adequate if the SPL difference reached at least 10 dB at 160 Hz or 8 dB at 250 Hz. When participants exhibited strong occlusion effects (indicated by visible physiological artifacts such as heartbeat in the waveform), fitting attempts were discontinued after two trials to prevent discomfort; regardless of the noise reduction level, a maximum of four attempts was made. There were four participants who could not reach 8 dB in the HO condition, even though participants subjectively felt the attenuation (and behaviorally could not hear the experimenter as well) and the sealing sensation in the ear canal. The occlusion effect was quantified by analyzing spectral differences between IEM and OEM while participants counted aloud at consistent volume. Initially performed only at setup and after earpiece adjustments, we later standardized these measurements to occur also after each noise condition, accounting for potential jaw-movement-induced ear canal deformations during speech that might displace the earpiece. Measurements were taken at 90, 160, 250, 500, 800, 1000, 1500, and 2000 Hz. Figure 3.2 shows the distribution of the noise reduction from the device fit check and its correlation with the occlusion effect at 250 Hz. Figure 3.2 also shows that LO data points (triangle) clustered on the left while HO data points (circle) spread toward the right, confirming that selected earplugs produced their intended category of occlusion effect.

3.3.1.4 Summary of the database

The comprehensive database contains speech recordings collected from 49 participants across 10 distinct acoustic conditions. For each condition, participants produced three types of speech

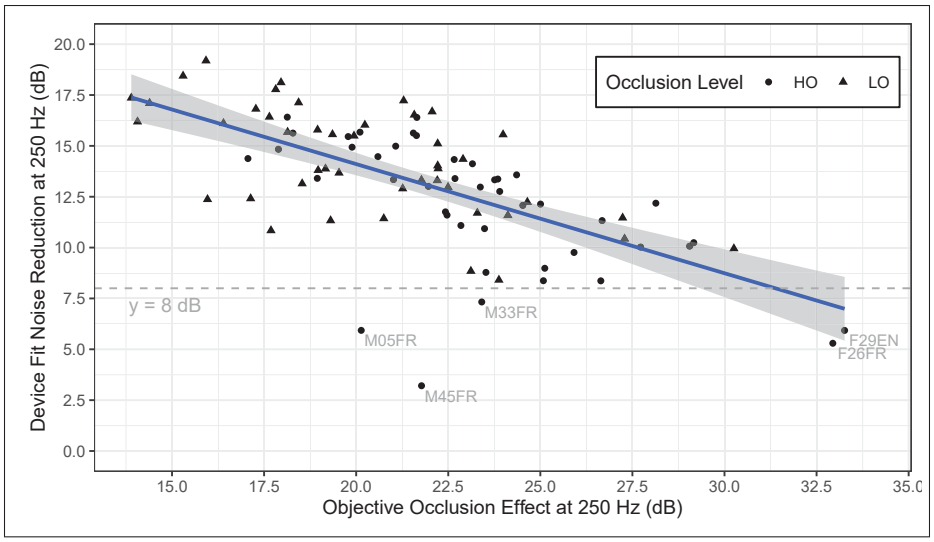


Figure 3.2 Correlation between noise reduction from device fit check and occlusion effect

X-axis shows the objective occlusion effect measured at 250 Hz, and y-axis shows the noise reduction from the device fit check. All the data point plotted here were taken right before the occlusion condition started, and each data point represents one participant in one occlusion condition (HO or LO), which is indicated by the shape of the data point. The traditional criterion at 8 dB is marked out, and the data points below 8 dB are marked out with their participants. A linear regression line was fitted, the shaded area showing its 95% CI.

samples: 20 phonetically balanced sentences from a standardized reading set, 15 sustained vowel productions (comprising 3 repetitions of 5 vowel types at 2.5 seconds each), and spontaneous descriptions of 3 different images lasting approximately 30 seconds per description. This structure yields about 10 minutes of recorded speech per condition, accumulating to 40 minutes per participant and totaling 1,960 minutes (32.7 hours) of speech across all participants.

While the OE condition used only the reference microphone in front of the mouth, all other conditions had five microphones simultaneously recording. This included the reference microphone, along with paired IEMs and OEMs. Notably, the IEM recordings were noise-free quality only during the NN conditions, while the reference microphone and the OEM remained uncontaminated by noise across all recordings.

The read sentences have been processed through Montreal Forced Aligner to establish word and phoneme-level alignment (McAuliffe, Socolof, Mihuc, Wagner & Sonderegger, 2017). Vowel productions share consistent timing markers across all recordings. For the picture description task, we first generated transcriptions using a robust speech-to-text model Whisper (Radford *et al.*, 2022) for a pseudo-truth, then we applied forced alignment with MFA.

The database also includes several valuable supplemental datasets: complete microphone calibration factors for signal conversion, participant demographic surveys, individual hearing profiles, and all the original recordings from the microphones in addition to the numeric measurements at specific frequencies from the occlusion effect measurements and device fit verification procedures.

3.3.2 Demonstrative and exploratory analysis with HIBiSCus

For the analysis in this paper, we used the sound pressure level measured from the phonetically balanced sentence reading task in front the mouth at 75 cm, hereafter referred to as speech level. The remaining speech tasks (vowel, picture description) with more acoustic features will be analyzed in future studies.

3.3.2.1 Data processing

Using the automated alignment timestamps, we removed silent segments before and after each utterance. For each participant and condition, we concatenated the processed sentences within each phonetically balanced set and calculated the overall A-weighted SPL using MATLAB R2022. We computed PTAs across the full frequency range (0.125-8 kHz) and categorized participants based on their hearing profiles in two approaches: (1) whether they had elevated (>15 dBHL) hearing thresholds at any frequency (any-frequency grouping); (2) whether they had elevated (>15 dBHL) PTA (PTA grouping). Our speech level analysis examined both SPL values and relative changes compared to the OE reference condition (ΔSPL).

3.3.2.2 Data analysis

In our analysis, we wanted to answer the following research questions:

- What is the effect of noise presence on speech level?
- Does variation in hearing thresholds explain variation in the Lombard effect?
- What is the effect of ear occlusion on speech level with and without noise presence?
- Does variation in hearing thresholds explain variation in the effect of occlusion on speech level?

To answer these questions, we used both descriptive statistics with data visualization with ggplot2 (Wickham, 2016) and statistical modeling with lme4 (Bates *et al.*, 2015) in R (Team, 2022).

For descriptive statistics, we calculated means and standard deviations (SDs) by acoustic condition across different hearing groups. We visualized speech level data across all acoustic conditions by participant and examined the relationship between PTA and ΔSPL by acoustic condition.

For statistical modeling, we used LME with a nested model comparison approach to test the statistical significance of fixed effects. All model comparisons used AIC and Bayesian Information Criterion (BIC) for model selection, with chi-square LRTs to assess statistical significance of model differences. Our modeling strategy addressed three specific goals in mind:

1) Assessing the comparability of OE and SE in quiet
To assess the similarity between speech level in the OE condition and SE in the NN condition, we fitted two models using data from OE and SE NN conditions:

Table 3.2 Model Equations for Comparing OE and SE

Model Type	**Equation**
Null model	SPL ~ 1
OE vs SE model	SPL ~ occlusion

Both models included random intercepts for participant and sentence set: (1 | participant) + (1 | set).

2) Examining the effects of noise, occlusion, and hearing grouping on speech level

Noise and occlusion were coded as *ordered* categorical variables, with noise levels (NN, LN, HN) and occlusion levels (SE, LO, HO), capturing the meaningful sequence from least to most intense level in both variables. This approach allows for the detection of linear, quadratic, and cubic trends across the ordered levels, with model coefficients interpreted as: Var.L (linear trend across the three levels), Var.Q (quadratic trend capturing acceleration or deceleration), and Var.C (cubic trend capturing more complex non-linear patterns).

We constructed five nested models with consistent random effects structure (by-participant intercepts with random slopes for occlusion and noise, plus sentence set intercepts: (1 + noise + occlusion | participant) + (1 | set)) while systematically varying the fixed effect structure:

Table 3.3 Model Equations for Investigating the Effects of Occlusion, Hearing Impairment Category, and Noise

Model Type	**Equation**
Full interaction model	SPL ~ occlusion * HI * noise
Partial interaction models	SPL ~ occlusion + HI * noise SPL ~ occlusion * HI + noise SPL ~ occlusion * noise + HI
Main effects only model	SPL ~ occlusion + noise + HI

We tested both hearing grouping methods (HI_AnyFrequency and HI_PTA) using this nested structure and used model comparison to identify the best-fitting approach. Lastly, we tested the statistical significance of each fixed effect term in the final model with further model comparison.

3) Examining the effects of noise, occlusion, and PTA on ΔSPL

For this analysis, we again used nested models with consistent random effects structure (by-participant intercepts with random slopes for occlusion and noise, plus sentence set intercepts: (1 + noise + occlusion | participant) + (1 | set)). Based on preliminary data exploration revealing a

potential breakpoint at PTA = 15 dBHL, we fitted piecewise regression models with the ordered categorical variables to test various interaction patterns:

Table 3.4 Model Equations for Investigating the Effects of Occlusion, PTA, and Noise

Model Type	**Equation**
Three-way interaction with breakpoint	$\Delta SPL \sim occlusion * PTA * noise + occlusion * PTA\_15 * noise$
Three-way interaction without breakpoint	$\Delta SPL \sim occlusion * PTA * noise$
Two-way interactions with breakpoint	$\Delta SPL \sim occlusion + PTA * noise + PTA\_15 * noise$
Two-way interaction without breakpoint	$\Delta SPL \sim occlusion + PTA * noise$ $\Delta SPL \sim noise + PTA * occlusion$

Through systematic model comparison, we tested whether the breakpoint at PTA = 15 dBHL significantly improved model fit and determined the nature of any interactions (with noise, occlusion, or both).

3.4 Results

3.4.1 Descriptive statistics

In Figure 3.3, the descriptive statistics show that speech level increased with noise level, with a maximum of 9.7 dB difference across noise conditions compared to a maximum of 1.5 dB difference across occlusion conditions. The effect of noise was more consistent and unidirectional, and overall, SDs increased with noise level. The effect of occlusion was smaller in magnitude and more variable. For the normal hearing group, participants spoke approximately 0.5 dB quieter in SE compared to OE conditions under no noise. The LO condition consistently increased speech level relative to both OE and SE across all noise levels. In the HO condition, speech level was similar to OE level in quiet but increased to LO levels when noise was present, with a slight 0.4 dB decrease only in high noise conditions. For the hearing impairment group

using PTA grouping, occlusion effects were minimal, with less than 0.4 dB difference between any adjacent occlusion conditions. The hearing impaired group using any-frequency grouping showed intermediate occlusion effects, which is larger than the PTA hearing impaired group but smaller than the normal hearing groups.

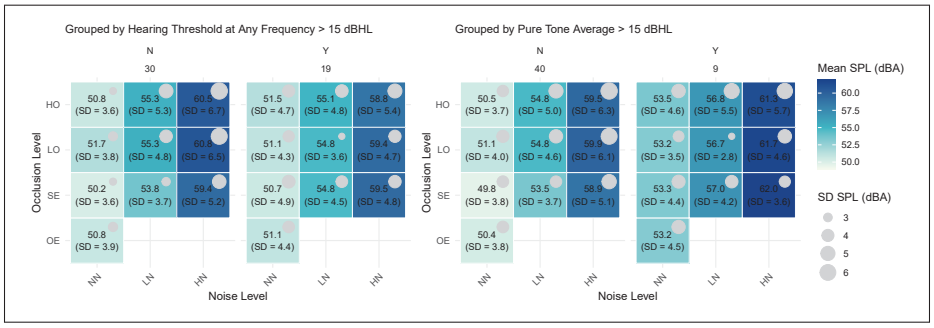


Figure 3.3 Mean speech level by acoustic condition and hearing group

Heat maps display mean speech level values with SDs in parentheses across three noise levels (x-axis) and four occlusion levels (y-axis). The left panel shows results using the any-frequency grouping, while the right panel presents results using the PTA grouping. Sample sizes are indicated as N (does not meet criteria) or Y (meets criteria) for each grouping. Color intensity represents mean speech level magnitude, and the size of the dot on the top-right corner represents the size of the SD.

Comparing the hearing groups, the hearing impaired groups showed greater variability in no noise conditions but reduced variability in high noise conditions. In addition, the two hearing grouping methods showed different patterns. Using any-frequency grouping, participants with hearing impairement spoke more quietly than those with normal hearing, but only under HN conditions when ears are occluded. Using PTA grouping, participants with hearing impaired consistently spoke louder across all noise conditions. These observations require statistical testing to determine significance, which is presented in subsequent analyses. Complete descriptive statistics with ranges are provided in the Supplementary Material.

In Figure 3.4, individual participant data showed large inter-subject variability in the effects of noise conditions on SPL. Consistent with the group-average findings, all participants

demonstrated increased speech level with the increase of noise, shown by the upward progression of curves from NN to HN. However, the magnitude of the reaction to noise varied greatly between participants, as shown by the gaps between those curves. Even within the any-frequency normal hearing group, some participants spoke as loudly in HN as others spoke in NN. Some participants' reaction to HN was equivalent to others' reaction to LN, despite having similar baseline OE levels.

The effect of occlusion showed even larger differences across participants. Recall that the overall average showed a down-up-down pattern (OE→SE→LO→HO). Only approximately one-third of participants followed this pattern, whose magnitude varied considerably. In addition, many participants deviated from this trend, showing partial adherence, alternative patterns, or even opposite trends.

Despite this individual variability in the effect of occlusion, there was consistency within participants across noise conditions. Most participants had overall similar responses to different occlusion levels across noise levels, as shown by parallel curve shapes that maintained similar patterns. However, certain participants showed large condition-dependent variations, with markedly different occlusion responses depending on the noise level.

Overall, no clear relationship between hearing status and speech level patterns was apparent from visual inspection of the individual curves, especially in the context of the large individual variability.

In Figure 3.5, we plotted PTA as a continuous variable in relation to ΔSPL. Using ΔSPL helps control for individual baseline differences shown in Figure 3.5 and the overall difference in hearing impaired group by PTA grouping. The LOESS curves revealed a notable shift in the relationship at PTA = 15 dBHL. For participants with PTA < 15 dBHL, the trend remained relatively flat or slightly positive. However, for participants with PTA > 15 dBHL, the relationship reversed – higher PTA were associated with smaller ΔSPL, with this negative trend being more pronounced in the two occluded ear conditions. Due to the relatively small sample sizes at PTA

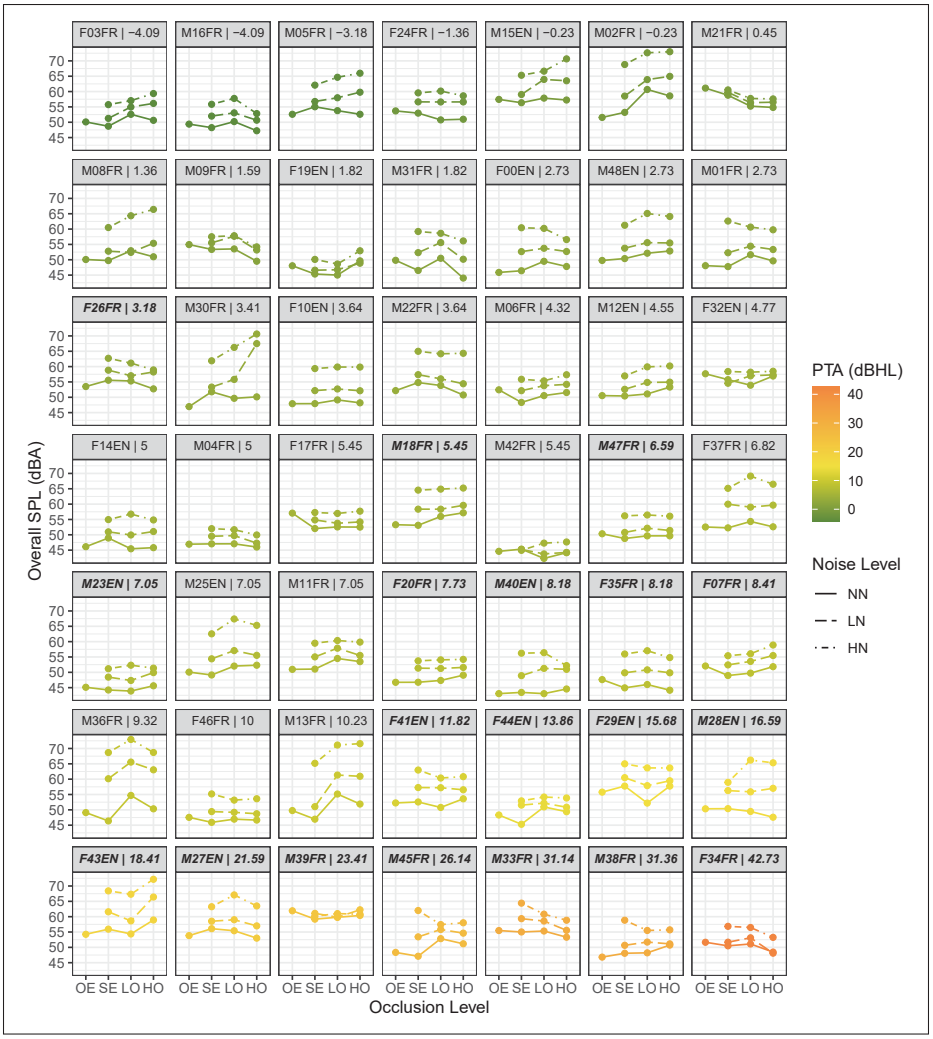


Figure 3.4 Individual participant speech level by acoustic condition

Each of the 49 subplots displays one participant's speech level (dBA) across the four occlusion levels (x-axis), with the three noise levels differentiated by line type. Participant ID and PTA are shown above each subplot. Participants with at least one frequency threshold above 15 dBHL are indicated in bold italic text. Line color represents PTA values.

> 15 dBHL, the CIs were much larger in this segment. Statistical modeling to further examine these trends is presented in subsequent analyses.

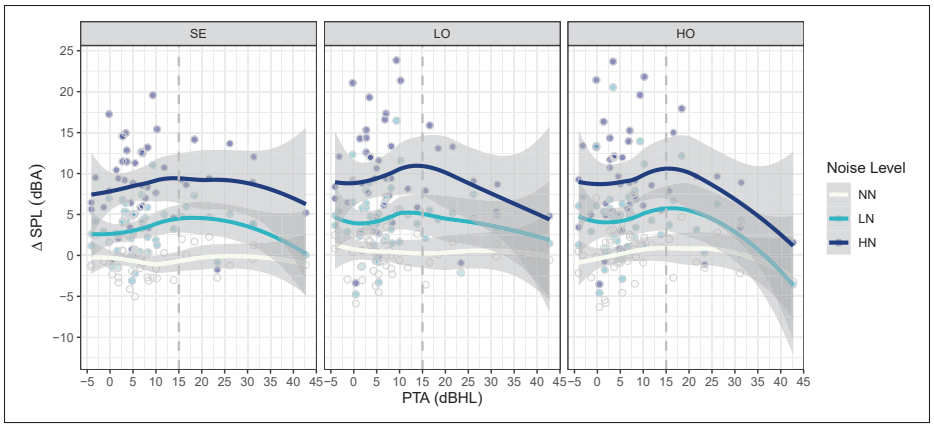


Figure 3.5 Relationship between PTA and ΔSPL by acoustic condition

Scatter plots show the relationship between PTA (x-axis) and ΔSPL (y-axis), where ΔSPL is calculated as the SPL difference from the baseline OE condition. Data are organized by occlusion levels (shown on top) with noise levels indicated by the color. Locally estimated scatterplot smoothing (LOESS) curves with 95% CIs (shaded areas) illustrate trends within each acoustic condition. The grey vertical line marks PTA = 15 dBHL.

3.4.2 Statistical modeling

3.4.2.1 Comparison of open ear and simulated open ear conditions in quiet

To assess the actual similarity between the SE condition in NN and the OE condition, we fit a LME with occlusion (OE vs. SE) as a fixed effect and random intercepts for participant and sentence set. The model estimate showed that participants spoke 0.29 dB quieter in SE than OE (95% CI [-0.72, 0.11]), but this difference was not significant (p = 0.146). Model comparison against an intercept-only model confirmed that including the occlusion effect did not significantly improve model fit (p = 0.153).

3.4.2.2 The effects of noise, ear occlusion, and hearing grouping

Here, we examined how noise level, occlusion level, and hearing status influence SPL. We also compared our two hearing grouping methods (any-frequency grouping and PTA grouping) to see which one better explains the overall speech level variations.

Model selection For the PTA grouping approach, model comparison using AIC and BIC revealed that the occlusion × HI + noise model (mod.ocxhi) and main-effects-only model (mod.main) provided the best model fits among the five nested models tested. While mod.ocxhi yielded a slightly lower AIC than mod.main, the chi-square LRT showed only borderline statistical significance (p = 0.088). Examination of the interaction effect coefficients in mod.ocxhi revealed non-significant effects (p = 0.2, p = 0.17). Subsequent model comparisons confirmed that all main effects (occlusion, noise, and HI) were statistically significant. Based on these results, we selected mod.main as our final model for the PTA grouping approach. We repeated the same nested model comparison procedure for the any-frequency grouping. Results showed that mod.ocxhi and mod.main again provided the best fits, with a slightly less significant difference between them (p = 0.1). Consistent with our approach for the PTA grouping models, we selected mod.main as the final model for the any-frequency grouping approach.

Direct comparison between the best-fitting models from the two grouping approaches demonstrated that the PTA grouping model provided superior fit, with lower AIC and BIC values. Therefore, our final model is the one with the PTA grouping HI variable:
SPL ~ noise + occlusion + HI_PTA + (1 + noise + occlusion | participant) + (1 | set). The complete model output summary is provided in the Supplementary Material.

Noise Noise demonstrated a significant overall effect with a predominantly linear pattern. The linear component was highly significant (noise.L: β = 6.25, standard error (se) = 0.41, p < 0.001), with the quadratic component also reaching significance (noise.Q: β = 0.47, se = 0.14, p = 0.002). Estimated marginal means revealed a progressive increase in speech level with increasing noise levels: NN = 51.97 dB (95% CI: [50.60, 53.34]), LN = 55.82 dB (95% CI: [54.33, 57.30]), and HN = 60.81 dB (95% CI: [59.01, 62.60]). Notably, participants increased

their speech levels more when moving from LN to HN (4.99 dB increase) than when moving from NN to LN (3.85 dB increase). The ses also increased with noise level (NN: 0.68, LN: 0.74, HN: 0.90).

Occlusion Ear occlusion showed a significant overall effect ($p < 0.05$), with the data better described by a quadratic rather than linear relationship across the three occlusion levels (occlusion.Q: β = -0.49, se = 0.20, p = 0.017). The linear component was marginally significant (occlusion.L: β = 0.50, se = 0.25, p = 0.052). Estimated marginal means of the main effect showed that SE produced the lowest speech level (55.64 dB, 95% CI: [54.24, 57.05]), while LO yielded the highest levels (56.60 dB, 95% CI: [55.05, 58.14]), and HO fell between these values (56.35 dB, 95% CI: [54.74, 57.96]). The overlapping CIs between LO and HO conditions confirm their similar SPL, with HO being slightly lower and more variable. Overall, ses increased progressively across occlusion conditions (SE: 0.70, LO: 0.77, HO: 0.80).

Hearing grouping PTA grouping was significantly associated with overall speech level (β = 3.11, se = 1.29, p = 0.020). Specifically, participants with PTA > 15 dBHL produced, on average, 3.11 dB higher speech level than those with PTA < 15 dBHL. Estimated marginal means were 57.75 dB (95% CI: [55.18, 60.33]) for the PTA > 15 dBHL group and 54.64 dB (95% CI: [53.35, 55.93]) for the PTA < 15 dBHL group.

Figure 3.6 provides a visual summary of the fixed effects of the best-fit model using the emmeans package (Lenth, 2022).

Random effects The random effects revealed substantial variability in speech level across participants. There was large variance in the by-participant intercept (σ^2 = 16.62). Additionally, participants varied in their reactions to both acoustic manipulations; the variance for slopes was 7.46 for noise and 2.46 for occlusion. Random effects showed moderate correlations between participant intercepts and slopes for occlusion (r = .42) and noise (r = .52). The slope terms for occlusion and noise were also moderately correlated (r = .53). The set-level variance was 0.07, suggesting minimal variability attributable to set. The residual variance was 1.80, indicating

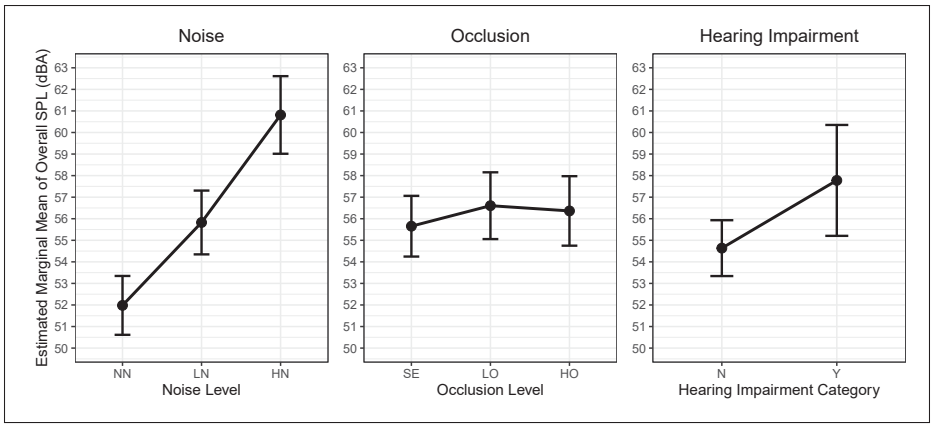


Figure 3.6 Effects of noise, occlusion, and hearing impairment on estimated marginal means of overall SPL

Each subplot displays the effect of one predictor shown on top on the estimated marginal mean of overall speech level, adjusted for the other two variables. Points represent the model-derived estimated marginal means, with error bars indicating 95% CIs. Lines connect the estimated marginal means to illustrate the pattern of the fixed effects across ordered factor levels.

that after accounting for variability due to participants and their responses to conditions, there was relatively little unexplained variability in speech level.

3.4.2.3 The effects of noise, ear occlusion, and PTA

Here, we examined how noise level, occlusion level, and PTA influence the speech level response. We used ΔSPL as the dependent variable, baselined by every participant's OE speech level. We also included a breakpoint at PTA = 15 dBHL to test whether the piecewise regression provides a better fit than a model without the breakpoint.

Model selection We first examined the inclusion of a breakpoint at PTA = 15 across the three interaction structures (three-way, two-way with noise x PTA + occlusion, two-way with occlusion x PTA + noise). Overall, there was a tendency for models including the piecewise breakpoint at PTA = 15 to show improved fit compared to models without it, although none of these improvements reached conventional levels of statistical significance. The three-way

interaction model showed a marginal improvement (p = 0.095), and the two two-way interaction models also exhibited similar trends (noise × PTA: p = 0.088; occlusion × PTA: p = 0.138). Model comparison among the three piecewise models showed that the full three-way interaction model did not significantly improve fit compared to the simpler two-way interaction models (Occlusion x PTA: AIC = 1939.0 and Noise x PTA: 1940.0, both with 28 parameters). The two two-way models themselves differed minimally in fit (ΔAIC, ΔBIC = 1.0), indicating they perform similarly in explaining the data. Complete model output summaries from both models are provided in the Appendix.

PTA × Noise The main effects of noise and occlusion aligned with results from the hearing group analyses. While PTA-related effects did not reach conventional significance (all p > 0.11), their estimates were accompanied by relatively wide CIs crossing zero. The $PTA_{<15}$ × noise linear interaction (β = 0.11, se = 0.07) had a 95% CI spanning [–0.03, 0.25], whereas the $PTA_{>15}$ × noise.L interaction showed an estimate in the opposite direction (β = –0.20, se = 0.13) with a 95% CI spanning [–0.45, 0.05]. Figure 3.7 presents the predicted lines from this model alongside the original data points.

PTA × Occlusion Again, the main effects of noise and occlusion aligned with results from the hearing group analyses. PTA-related effects were more nuanced here. The $PTA_{<15}$ × occlusion.L interaction was weak and non-significant (β = 0.02, se = 0.05; p = 0.66), but the $PTA_{<15}$ × occlusion.Q interaction was borderline (β = 0.07, se = 0.04; 95%CI: [-0.00, 0.15]; p = 0.06), suggesting a potential modulation of this nonlinearity by PTA. Furthermore, the occlusion.Q × $PTA_{>15}$ interaction was statistically significant (β = -0.15, se = 0.07; 95%CI: [-0.29, -0.01]; p = 0.04), indicating a small but rather reliable opposing pattern above 15 dBHL. Figure 3.8 presents the predicted lines from this model alongside the original data points.

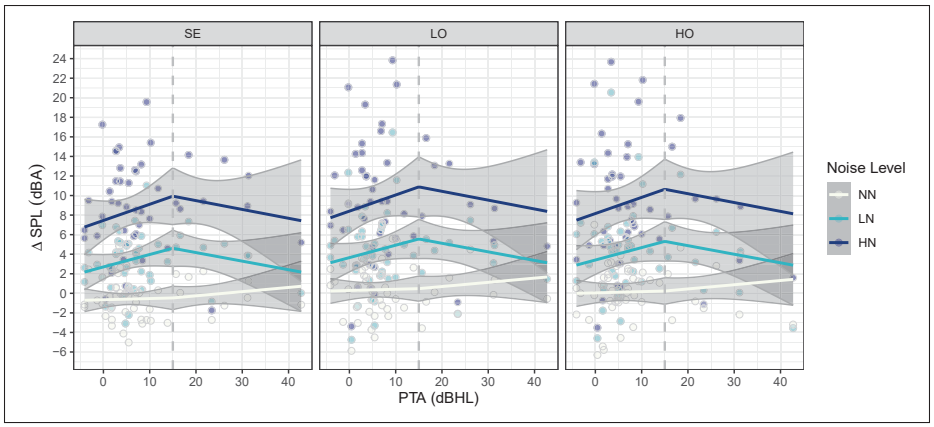


Figure 3.7 Model-predicted relationship between PTA and ΔSPL from the piecewise PTA-noise interaction model

Scatter plots show the relationship between PTA (x-axis) and ΔSPL (y-axis), where ΔSPL represents the speech level difference relative to the baseline OE condition. Data are grouped by occlusion levels (shown across panels), with noise levels indicated by color. Overlaid lines represent model predictions from the piecewise PTA-noise interaction model. Shaded ribbons indicate 95% CIs around the model-predicted lines. The grey vertical line marks PTA = 15 dBHL, the breakpoint used in the piecewise regression.

3.5 Discussion

3.5.1 Effects of noise

Our results show that noise clearly affects speech level, which matches what we know about the Lombard effect. Both the descriptive statistics and the statistical models confirmed this effect, and it was very strong ($p < 0.001$).

When we compare the LN and HN conditions (a 15 dB difference), speech level increased by about 5 dB on average. This means speech level went up by 0.33 dB for every 1 dB increase in noise. When we compare the NN and LN conditions, the difference between the two was smaller at 3.85 dB. In NN, the experiment room was very quiet (< 20 dBA), while the LN condition played a noise at 70 dBA. That's a much bigger difference than the 15 dB step between LN and HN, even if we consider the finding in the literature that Lombard effect would only be

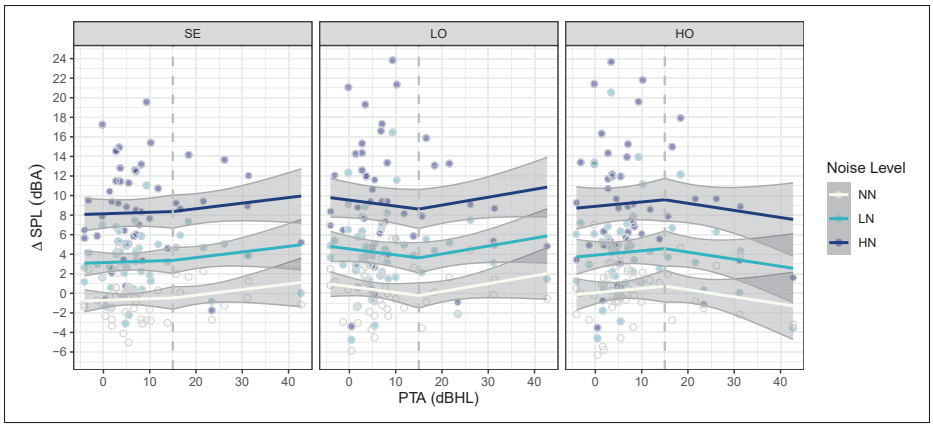


Figure 3.8 Model-predicted relationship between PTA and ΔSPL from the piecewise PTA-occlusion interaction model

Scatter plots show the relationship between PTA (x-axis) and ΔSPL (y-axis), where ΔSPL represents the speech level difference relative to the baseline OE condition. Data are grouped by occlusion levels (shown across panels), with noise levels indicated by color. Overlaid lines represent model predictions from the piecewise PTA-occlusion interaction model. Shaded ribbons indicate 95% CIs around the model-predicted lines. The grey vertical line marks PTA = 15 dBHL, the breakpoint used in the piecewise regression.

activated when noise is above 43.3 dBA. This means that speech level went up by 0.15 dB for every 1 dB increase in noise. Previous studies have reported Lombard effect magnitude at 0.2 ∼ 1 dB/dB as mentioned in Section 3.2.2, and we can see that the effect magnitude we have in our current study fall on the lower end or even lower than the already reported numbers. Several factors may explain this smaller effect size. First, previous research has shown that artificial noise stimuli (like those used in our study) tend to evoke weaker Lombard effects than natural noise environments (Giguère *et al.*, 2006). Second, the presence of ear occlusion may have influenced the results - we examine this effect more closely in the interaction effect subsection in Section 3.5.3.

We saw considerable individual variability in responses to noise, as evident in Figure 3.4. Both descriptive statistics and statistical modeling showed that this variability increased with higher noise levels. The random slope of noise by participant in the mixed-effects model(s) revealed substantial between-participant differences in noise adaptation. Additionally, we

found a moderate positive correlation between participants' baseline speech levels and their noise-induced voice amplification, which indicates that those who spoke louder at baseline tended to show greater increases in noise. Future studies can examine whether this correlation persists in other tasks from our database, which helps determine if it reflects consistent individual differences.

3.5.2 Effects of ear occlusion

Our comparison of OE and SE conditions showed an average 0.4 dB decrease in speech level for SE. While this small effect is both statistically non-significant ($p = 0.146$) and likely inaudible, the individual data tell a more complete story. Looking at each participant's response in Figure 3.4, we see people reacted differently—compared to OE, some spoke quieter in SE while others spoke louder or had minimal change. From Figure 3.5, we can also see the individual data points of this if we look at the NN points in the SE condition, and the range of difference was -5.0 ~ 4.8 dB. This explains why the average effect was small, yet the CI was large: these counteracting individual trends were likely combined in the analysis.

The simulated open-ear condition, despite avoiding ear-canal occlusion, still produces systematic though opposing effects on speech level control. The coexistence of positive and negative responses implies multiple underlying mechanisms may be at work simultaneously. This likely stems from how the open-back headphones interacted with the pinna's natural acoustics. Their physical presence around the outer ear may have altered self-perception of speech level in two competing ways: the dome-like ear cups could (1) selectively amplify certain frequencies through resonance, making participants perceive their voice as louder, or (2) dampen high-frequency components that are normally enhanced by pinna reflection, creating perceived loudness reduction. This dual mechanism could potentially explain why there were two opposing trends. Future work using probe-microphone could quantify these acoustic changes.

Both occluded conditions (LO and HO) overall led to louder speech compared to SE, but the HO condition did not lead to higher speech level than in LO. This general pattern can likely

be explained by this: the overall attenuation of the AC component of the self-perceived voice led to increased speech level, while the amplified BC component may become uncomfortably prominent and “boomy", leading speakers to lower their speech level for comfort. In addition, the silicon flange earplugs (used in HO) have been shown to attenuate external sounds less than the foam earplugs (used in LO) (Tufts & Frank, 2003). This means that the AC path speech was less attenuated in HO than in LO, which could also lead to less increase in speech level in HO compared to LO. However, there were again substantial individual variability evident in both the random effects results and individual response patterns. This is not surprising because again we have two opposing forces: attenuated AC path and amplified BC path, and it would not be surprising if some prioritizes one cue over another while others the reversed, leading to diverging responses. For example, some participants spoke louder in the HO condition, likely because they were compensating for the further-reduced audibility of their AC voice component—since it has become even more masked by the stronger BC amplification in HO—while either tolerating or not experiencing discomfort from the amplified BC component.

The observed effects likely depend on both the earplug’s attenuation properties and the degree of occlusion, which varies across individuals due to differences in ear canal anatomy and earpiece fit. While earplug types in this database provided a categorical grouping, examining individual device fit characteristics and occlusion effect measured by IEMs and OEMs could potentially better explain the nuanced patterns of speech level differences, and we may even better be able to approximate the AC and BC component.

Regarding occlusion effect measurement, Figure 3.2 demonstrates the correlation between noise reduction from device fit checks and objective occlusion effects. Existing device fit check procedures do not account for this correlation, which may lead to over-conservative assessment that cannot be resolved. Adding a baseline measurement of occlusion effect in silence (without additional sound sources) during device fitting could address this limitation.

Future studies can also investigate if the degree of communicativeness of the speech tasks affect the effect of occlusion on speech production, revealing whether speakers increasingly rely on

AC feedback when prioritizing more listener-directed communication (picture description vs sustained vowel), as the AC component better represents how voices are perceived externally. Frequency-specific analyses (e.g., high vowel production or fricative production) may further clarify how speakers adjust strategies in response to the occlusion-induced changes, as the changes from AC and BC target different frequency components.

3.5.3 Potential interaction effects between noise and occlusion

Although our model comparisons did not support including a noise-by-occlusion interaction term, several observations suggest this interaction could be something to research more on. First, the effects of noise and occlusion operate on different scales and may interact in complex ways that our current sample size lacks power to detect. Individual response patterns show cases where participants' reactions to combined noise and occlusion deviate from simple additive effects—some exhibit parallel increases while others show divergent trends. An interaction between noise and occlusion is plausible because these factors influence speech in competing ways. When there is noise, people naturally speak louder because of the disruption in the AC component. Adding ear occlusion makes this more complicated: a stronger occlusion effect means hearing oneself more clearly through BC, while simultaneously hearing yourself less clearly through AC due to attenuation from the earplugs and BC masking, yet what the strategy is may differ from one to another. This variability with potentially opposing effects could explain again why our overall model did not detect a significant interaction effect. In addition, the moderate correlation in the random effects between participant-specific noise and occlusion slopes further hints that individuals who react strongly to one acoustic manipulation may react strongly to the other, though this speaks more to between-participant variability than a simple interaction. Overall, future studies should pay more attention to individual variability by further investigating cases where interaction effects do appear. It would be interesting to focus on exploring what explains these difference—whether particular compensation strategies are being used—by potentially examining phoneme-level differences or comparing responses across different speech tasks.

3.5.4 Effects of hearing thresholds

We found that PTA grouping explained speech level differences better than any-frequency grouping, which is understandable as having only one or a few frequencies at above 15 dBHL likely does not change the perception of the noise level drastically, especially in the context of the large individual differences observed regardless of hearing thresholds. This could also have come from a lack of the necessary statistical power to detect the more subtle yet widely variable differences.

The results from the PTA grouping model revealed rather unexpected patterns. While it would have made sense if people with higher hearing thresholds reacted less to noise because they perceived it less, we in fact did not observe a statistically significant interaction effect between hearing grouping and noise condition in this approach; rather, on average, people with hearing impaired in our sample spoke louder regardless of the acoustic conditions. Rather than concluding that there is no interaction effect, a more plausible interpretation is that on average individuals with hearing impairment may have developed a habitual tendency to speak louder, and this elevation in overall speech level could have outweighed the potential effect of reduction in the Lombard effect, particularly given the substantial variability observed even within the normal-hearing group.

While the categorical grouping approach showed that individuals with higher hearing thresholds tend to speak louder overall, our continuous modeling of PTA was aimed at going beyond group-level averages and targeting the participants' responses to the acoustic manipulations by removing the confounding influence of overall speech level differences across hearing groups. This approach allowed nonlinear patterns to emerge organically from the data: both the LOESS-smoothed empirical curves and the model predictions revealed a turning point near PTA = 15 dBHL, which aligns with the conventional boundary for normal hearing. While the change did not reach conventional levels of statistical significance—likely due to limited statistical power—the pattern is nonetheless compelling, suggesting that individuals with PTA above 15 dBHL may react differently to the acoustic manipulations. In this context, statistical

non-significance should not be equated with a lack of meaningful trends, especially when the patterns observed are theoretically coherent and emerge consistently across multiple analytical approaches.

This trend was especially evident in the PTA × noise model, where we observed a positive slope before 15 dBHL and a negative slope afterward, across all occlusion conditions. In contrast, the PTA × occlusion model showed more differentiated patterns across occlusion types: under the SE condition, the slope remained upward; under the LO condition, the slope flipped from negative to positive at 15 dBHL; and under the HO condition, the trend was flat or slightly positive before 15 dBHL and declined afterward. Interestingly, the HO condition showed consistent behavior across both models, possibly indicating an interaction between hearing level and occlusion not captured elsewhere.

With only nine participants completing a 3 (noise) × 3 (occlusion) within-subject design, the sample size was likely underpowered to detect two-way interactions (such as PTA × noise or PTA × occlusion) and even less so for potential three-way interactions. In fact, in our two-way interaction models, several CIs narrowly crossed zero, suggesting that the observed effects may be meaningful but require further investigation. Thus, we caution against overinterpreting these trends, whether too conservatively by dismissing them outright, or too liberally by overstating their significance.

Additionally, while the directionality is suggested, the slope differences were generally modest, typically less than 0.5 dB change in SPL per 1 dB change in PTA. This raises an important question: are these subtle changes meaningful enough to serve as reliable indicators of hearing status? The answer may depend on inter-individual variability, which was substantial in our dataset: some participants with normal hearing showed minimal Lombard effects, while others with hearing impairment showed large Lombard effects. Thus, gathering current evidence from the descriptive statistics, statistical modeling and the individual response curves, it does not seem that the speech level and Lombard effect magnitude alone are sensitive enough to distinguish hearing status. However, tracking these relationships longitudinally may reveal more meaningful

patterns within individuals, taking into consideration the “baseline" magnitude of their responses to noise.

One potential direction for future research involves refining how PTA is quantified. In this study, we used an average PTA across frequencies ranging from 0.125 to 8 kHz. However, there are different ways to calculate PTA, and it may be worthwhile to explore whether specific frequency regions are more predictive of differences in speech level modulation. Another important direction is to expand beyond overall speech level as the primary acoustic measure. Future work could explore more detailed phoneme-level analyses (as already mentioned above) and examine additional acoustic features such as pitch, formants, spectral center of gravity, and other parameters. This would allow for a more comprehensive use of the rich dataset, especially given the diversity of speech tasks included in the study.

3.6 Conclusion

This study introduced a comprehensive speech production database with varying listening conditions available by request. We demonstrated its utility for investigating complex auditory-based speech control phenomena using speech level as an example. Our analysis revealed that speech level is affected by noise, occlusion, and hearing impairment and it exhibits considerable individual variability that challenges traditional group-level interpretations.

The theoretical implications are significant. Rather than simple linear relationships, we found that the effect of ear occlusion on speech production is not unidirectional, with substantial individual differences suggesting that speakers may be using different auditory feedback control strategies. Preliminary investigation on hearing impairment showed a categorical shift around PTA = 15 dBHL, where individuals with greater hearing impairment became less reactive to noise especially in high occlusion condition, while those without impairment showed increased response with the increase of PTA. These nonlinearities in occlusion and hearing impairment may explain conflicting results in previous literature, which assumed linear relationships. Future research should also investigate potential noise-occlusion interactions at the individual level.

The categorical versus continuous nature of hearing impairment effects also requires further investigation with larger datasets.

From a practical perspective, longitudinal tracking of one's speech level patterns could potentially serve as an early indicator for hearing impairment detection before clinical thresholds are reached, but more research is needed to develop algorithms sensitive enough for this application, ideally also exploring other acoustic parameters. Additionally, we showed that objective occlusion effect correlates with device fit check noise reduction measurement and it suggests that current fitting procedures could be improved by incorporating individual occlusion baseline.

3.7 Author Declarations

The authors have no conflicts of interest to disclose. The data of this study are available upon request. This research project received ethics approval from the ethics committee of the École de technologie supérieure (H20211107).

3.8 Acknowledgments

This research was supported by Natural Sciences and Engineering Research Council of Canada (Discovery grant: RGPIN-2021-03182), Fonds de recherche du Québec's Master's training scholarship, CIRMMT, and the École de technologie supérieure's Marcelle Gauvreau Engineering Research Chair in Multimodal Health Monitoring and Early Disease Detection with Hearables.

CHAPTER 4

THE EFFECTS OF MICROPHONE POSITIONING IN HEARABLES ON VOICE QUALITY AND F0 MEASUREMENTS

Xinyi Zhang[1,2] , Arian Shamei[1,2] , Alessandro Braga[1,2] , Rachel Bouserhal[1,2]

[1] Department of Electrical Engineering, École de technologie supérieure, Montréal, Québec H3C 1K3 Canada

[2] Centre for Interdisciplinary Research in Music Media and Technology, Montréal, Québec H3A 1E3 Canada

Paper published in “The Journal of the Acoustical Society of America" on September 1, 2025.

4.1 Abstract

Voice quality and fundamental frequency metrics are important indicators of motor function and hold promise for health monitoring. Recent advances in hearables have enabled the longitudinal monitoring of speech production and its changes. Hearables can record speech from IEM and OEM, but it remains unclear how these measurements from hearables compare to the laboratory gold standard, a microphone placed in front of the mouth. This study examines voice quality and F0 measurements across the IEM, OEM, and the standard method (REF) using parallel recordings. Results showed that the IEM introduced more variability overall; increases in jitter, HNR and F0 maximum and standard deviation and decreases in F0 minimum were seen for females. Decreased shimmer and increased HNR were seen in the OEM. The causes of these differences were discussed. The findings indicate that the hearable-based measurements may not align with REF standards, suggesting the need for new standards specific to hearables. Preliminary observations of sex-based differences require further investigation with adequately powered and balanced samples to determine their significance and generalizability. Future research should further explore factors such as occlusion effect and sex-specific differences (e.g, F0 range) in the relationship between hearables and REF measurements.

4.2 Introduction

Recent advancements in the use of microphones placed within the ear canals (in-ear microphone, or IEM for short) have opened new possibilities for the acoustic evaluation of human biomarkers, especially when the IEM is pointed inward within an occluded ear canal. Hearables, wearables that are located in and around the ear, provide a window to a plethora of human bodily signals, such as heart rate, breathing, swallowing, eye movement, and speech (Mehrban *et al.*, 2024; Chabot *et al.*, 2021; Röddiger *et al.*, 2022; Goverdovsky *et al.*, 2017). The continuous long-term monitoring of speech provides invaluable insights into evaluating health and disease progression, leveraging the fact that speech production is a complex biological process involving multiple physiological systems (Botelho, Abad, Schultz & Trancoso, 2024).

For example, neurodegenerative disorders such as Alzheimer's disease (AD) and Parkinson's disease (PD) exhibit reduced tongue movement, resulting in altered formants and a smaller acoustic vowel space than those of healthy controls (Shamei, Liu & Gick, 2023; Skodda, Visser & Schlegel, 2011). Accordingly, changes in muscle tone and laryngeal motor control in both diseases alter fundamental frequency and measurements of voice quality such as jitter, shimmer, and noise-to-harmonics ratio (Jiménez-Jiménez *et al.*, 1997; Ebbutt, Shamei, Purnomo & Gick, 2021). Consequently, speech can serve as a significant indicator of an individual's health or disease state. Measurements of voice quality are responsive to treatment, for example elevated measurements of jitter and shimmer in PD decrease in response to dopaminergic drugs (Azadi, Akbarzadeh-T, Shoeibi, Kobravi *et al.*, 2021), which suggest that they accurately represent the severity of disease symptoms. The sensitivity of voice-based measurements for the discrimination of PD has generated substantial interest in their use for early disease detection systems (Solana-Lavalle, Galán-Hernández & Rosas-Romero, 2020). While these perturbation measures may not always outperform more complex machine-learning acoustic features in all detection contexts, their clinical interpretability, computational efficiency, and established theoretical foundation make them valuable for continued investigation, particularly for practical clinical implementation.

Traditional methods of voice analysis often rely on microphones placed in front of the mouth, but the IEMs of hearables present a promising alternative for continuous long-term monitoring of voice quality. The placement of the microphone within the ear canal reduces exposure to background noise and extraneous speech, while maintaining continuous recording of the sounds produced by the wearer. However, the effects of human physiology on intracranial recordings must be taken into account; the bone and tissue of the skull act as a low-pass filter, attenuating frequencies above 2 kHz, and the occlusion of the ear canal amplifies signals in the lower frequencies (Bouserhal *et al.*, 2017b). These differences raise the question as to whether insights into acoustic speech changes as measured from traditional microphones can be applied directly to IEM speech. Additionally, hearables often have an OEM, which captures air-transmitted sound similar to traditional microphones. Here, we aim to directly compare the fidelity of IEM and OEM to traditional recording methods. We compare measurements of voice quality (jitter, shimmer, harmonics-noise ratio) and F0 (mean F0, variability of F0, min and max F0) across the IEM, OEM and standard microphone speech recordings taken from 8 speakers from the open-access SpEAR database (Bouserhal *et al.*, 2019).

4.2.1 Measurements of voice quality

The phonation process sets the vocal folds into vibration, producing quasi-periodic sound waves, particularly in vowels and other voiced sounds. Jitter and shimmer measure the variability during phonation, with jitter assessing the temporal variability of the F0 (i.e., the time between consecutive vocal fold vibrations) across time, and shimmer measuring the variability in amplitude across these pulses.

It is important to note that there is no single manner in which metrics of voice quality such as jitter and shimmer may be measured. They can be measured by comparing the time difference across any number of pulses or in sets of pre-determined length. F0-independent metrics (i.e. jitter, shimmer) are inherently time-sensitive and dependent on basic properties of F0 or speech rate, thus different evaluation methods can be employed for different languages and speaking environments. For example, a language with shorter syllables would benefit from a shorter

window of analysis, yet the number of F0 pulses within a syllable is influenced not only by the length of a syllable but also by the F0 of the speaker (with a higher F0 naturally resulting in a higher number of pulses within an identical timeframe).

The Harmonic-to-Noise Ratio (HNR) measures the proportion of harmonic sound to noise within a voiced signal and it is reported in dB. A higher HNR suggests a clearer voice, whereas a lower HNR indicates weak harmonics or aperiodic vibration which may sound breathy, hoarse, or creaky.

Previous research has investigated factors that may influence voice quality measurements. For example, Brockmann-Bauser, Bohlender & Mehta (2018) demonstrated that voice SPL negatively correlates with jitter and shimmer while positively correlating with HNR when participants were asked to speak in three levels of vocal intensity (soft, comfortable, and loud). The authors discussed that this could be due to an underlying physiological mechanism related to a stiffer and more stabilized vocal fold in higher vocal intensity. However, this work focused on physiological (and behavioral) factors affecting voice production itself. In contrast, our research examines how acoustical differences arising from microphone positioning relative to the same sound source may influence voice quality measurements—an area that has not been previously investigated.

4.2.2 Speech recorded by REF, OEM, IEM

In a laboratory setting, speech is typically recorded with the microphone placed directly in front of the mouth (Vogel & Morgan, 2009). This ensures that the microphone can capture the sound waves directly from the mouth, maintaining a highly accurate representation of the speech signal and minimizing distortion and the effect of indirect AC. The microphone ideally has high sensitivity, a flat frequency response, and a low noise floor (Švec & Granqvist, 2010). However, this "standard" is not necessarily achievable for miniature microphones used in hearables, and the placement of the microphone is not in front of the mouth. Thus, it is within expectations that speech recordings via wearables may differ from traditional microphones, hereon referred to as "reference" microphones (REF).

Speech signals recorded by the IEM in an occluded ear canal have been transmitted through the bone-and-tissue medium. This medium acts as a low-pass filter that attenuates frequency contents above 2 kHz. Due to the occlusion of the ear canal, the sound waves are trapped within the closed tube of the ear canal created by the earplug and the eardrum, which enhances what the IEM captures. This effect of occlusion (termed occlusion effect) can be objectively measured as the difference in sound pressure level between the IEM and the OEM when the wearer produces a sound, where the IEM is overall louder than the OEM. The maximum amplification is seen at lower frequencies. For example, Saint-Gaudens *et al.* (2022) reported ~20 dB of difference at 160 Hz and ~17 dB at 250 Hz. Speech recordings from IEM are often described as boomy and muffled (e.g., Bouserhal *et al.* 2019; Saint-Gaudens *et al.* 2022). OEM records only air-transmitted sounds like REF, but it is placed at the opening of the ear canal. This recording is less direct and is more prone to the effect of indirect AC in the environment.

4.2.3 Predictions

This is the first study to assess the impact of ear occlusion on the measurement of voice quality. We propose the following hypotheses regarding the potential influence of ear occlusion on IEM recorded measures.

For jitter, the occluded ear canal may produce more complex air vibrations compared to OEM and REF conditions. This increased complexity could lead to greater perturbation, resulting in higher variability in jitter measurements. We also anticipate a corresponding increase in the SD of the fundamental frequency, with minimal impact on the minimum and maximum F0, and no effect on the mean F0.

The amplification of lower frequencies due to occlusion may enhance the magnitude of these variations, making them more detectable by the microphone. This effect is particularly relevant for shimmer, as the increased amplification of lower frequencies might accentuate amplitude variations. Regarding the HNR, the amplification of lower frequencies—where most of the harmonic energy resides—could lead to an increase in HNR in IEM recordings.

For what concerns the comparison of OEM and REF conditions, we expect their differences to be minimal, especially when compared to their differences from the IEM condition.

Additionally, physiological differences between male and female subjects may also influence these measures, though the specific effects are difficult to predict given limited research on sex-based differences in occlusion. Three key factors may interact to create these differences. First, female subjects typically have a higher fundamental frequency than male subjects due to their shorter and smaller vocal folds (Gick, Schellenberg, Stavness & Taylor, 2019). Second, Yu *et al.* (2015) have shown that males typically have longer and wider ear canals. This anatomical difference may affect how the hearables fit within the ear canal, altering the degree of the OE. Third, since the occlusion effect amplifies lower frequencies in a frequency-dependent manner, these anatomical and vocal differences may combine to create sex-based variations in our measurements. However, because female's F0 values still fall within the frequency range affected by occlusion, and because ear canal shape differences may alter the occlusion pattern across frequency, we cannot make specific directional predictions about these effects at this stage.

4.3 Methodology

4.3.1 Dataset and processing

Speech recordings from eight native English speakers (5F, 3M) were obtained from the SpEAR database (Bouserhal *et al.*, 2019), where speakers were instructed to read Hearing-In-Noise Test sentences in an audiometric booth wearing a pair of hearables. The HINT sentences are short, phonetically balanced sentences using common vocabulary (e.g., "She spoke to her eldest son," "An oven door was open"). In terms of the microphones, the IEMs and the OEMs were Sonion 50GE31 and Knowles FG23652 respectively, and a 1-inch low-noise microphone GRAS 40HF (GRAS Sound & Vibration, Holte, Denmark) was used as REF. Further details regarding the database and the apparatus can be found in Bouserhal *et al.* 2019 where it was originally published. This audiometric booth (L * W * H: 364 cm * 279 cm * 200 cm) is certified for the

ANSI S12.6 standard (American National Standards Institute, 2008) and ISO 4869 standard (International Organization for Standardization, 1994) in terms of uniformity, directionality, and ambient noise. The ambient noise level is less than 20.5 dBA, and the reverberation level is low, with reverberation time based on 30 dB decay (RT30) less than 0.06s; this ensures the clarity of the speech recordings. For each speaker, parallel recordings from the left IEM (placed inside of the left occluded ear canal), left OEM (placed outside of the left ear canal), and the REF (placed 30 cm in front of the mouth) were selected to be used in the current study. The SNR calculated from the recordings from the IEM was 36.7 dB, the OEM 38.2 dB, and the REF 38.7 dB. All audio files were used with their original sampling rate (48 kHz) and bit-depth (32 bits).

We first used the Montreal Forced Aligner (McAuliffe *et al.*, 2017) to provide an automatic phonetic alignment for each recording. This procedure provides the starting and ending timestamps of each phoneme. Research assistants then manually verified and adjusted all automatic alignments to ensure accurate segmentation. The manual verification process used an acoustic-first approach, where trained assistants examined spectrograms and waveforms to identify segment boundaries, followed by auditory validation. In cases where coarticulation made boundaries ambiguous, adjustments were made to optimize for perceptually salient transition points that best distinguished adjacent segments. An example of our segmentation is shown in Figure 4.1. Using the audio analysis software Praat (Boersma & Weenink, 2001), measurements of jitter, shimmer, HNR, and F0 were extracted using a Praat script. F0 was calculated using Praat's 'To Pitch (ac)' function with default settings, except for pitch floor and ceiling values which were individually determined for each participant through manual examination to avoid pitch halving and doubling errors. Jitter, shimmer, and HNR were extracted using Praat's 'To PointProcess (periodic, cc)' function followed by the Voice Report function, using default parameters except for the participant-specific pitch floor and ceiling values. Since it remains unknown exactly how the different microphone placements may affect measurements of these metrics, we have opted to include variations of each metric. Table 4.1 provides a brief description of each selected variation and how it is calculated from Praat.

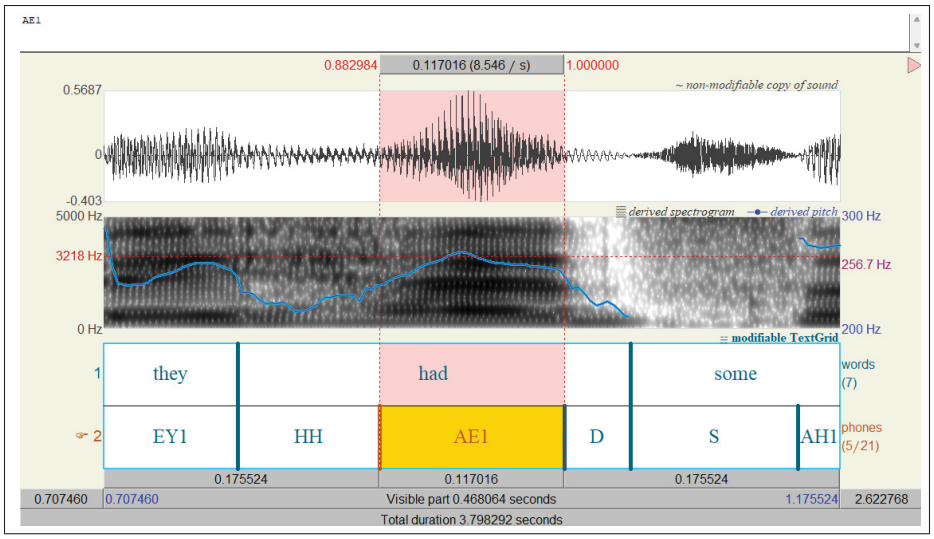


Figure 4.1 Example of phonetic segmentation

Table 4.1 All metrics used in the current study and how they are calculated

Metrics		**Definition**
Jitter	Local (%)	mean absolute difference in duration between consecutive periods divided by the mean period
	Local (μs)	mean absolute difference in duration between consecutive periods
	RAP (%)	Relative Average Perturbation: mean absolute duration difference between a period and the mean of this period and its two adjacent periods, divided by the mean period
	PPQ5 (%)	Five-Point Period Perturbation Quotient: mean absolute difference between a period and the mean of this period and its four closest neighbors, divided by the mean period
Shimmer	Local (%)	mean absolute difference between the amplitudes of consecutive periods divided by the mean amplitude
	Local (dB)	mean absolute log10 of the difference between the amplitudes of adjacent periods, multiplied by 20
	APQ3 (%)	Three-Point Amplitude Perturbation Quotient: mean absolute difference between the amplitude of a period and the mean of the amplitudes of its neighbors, divided by the mean amplitude.
	APQ5 (%)	Five-Point Amplitude Perturbation Quotient: the mean absolute difference between the amplitude of a period and the mean of the amplitudes of this period and its four closest neighbors, divided by the mean amplitude
	APQ11 (%)	Amplitude Perturbation Quotient: the mean absolute difference between the amplitude of a period and the mean of the amplitudes of this period and its ten closest neighbors, divided by the mean amplitude.
HNR	Mean (dB)	Ratio of energy deriving from harmonics of F0 relative to non-harmonic sounds
F0	Mean (Hz)	Mean F0 across the token
	SD (Hz)	Standard deviation of F0 across the token
	Min (Hz)	Lowest F0 measured across the token
	Max (Hz)	Highest F0 measured across the token

All the metrics above were extracted from the vowel segments identified within the TextGrid. Crucially, only stressed vowels were retained to ensure a sufficient length and quality for accurate measurements of each metric. Table 4.2 outlines the total number of vowel recordings used

in the present study for each speaker (mean = 994, sd = 47), removing mispronunciations and missing annotations.

Table 4.2 Number of vowel recordings per speaker

	F1ENG	F3ENG	F4ENG	F5ENG	F6ENG	M1ENG	M2ENG	M6ENG
REF #	343	326	319	336	362	321	314	329
IEM #	343	326	319	336	362	321	314	329
OEM #	343	326	319	336	362	321	314	329
Total #	1029	978	957	1008	1086	963	942	987

4.3.2 Data analysis method

4.3.2.1 Spectral content comparison of the three microphones

Using MATLAB R2022b (The MathWorks Inc., 2022), sound files were grouped by the participants' sex and the recording microphone, and files within each group were concatenated. Then, they were further converted from a representation in volts to sound pressure in Pa with microphones' sensitivity calibration factors. The reference microphone's sensitivity calibration was done with a standard microphone calibrator that plays a 1 kHz pure tone at 94 dB. We recorded this signal with the reference microphone and its sensitivity calibration factor was the scaler factor that needed to be applied to the recorded signal for it to be 94 dB at 1 kHz. To measure the miniature microphones' sensitivity calibration factors, white noise (SPL > 85dB) was played and recorded in a reverberation chamber simultaneously by the miniature microphones and a 1/2-inch reference microphone placed as close as possible to each other without touching. Then, by applying the reference microphone's sensitivity calibration factor (measured in the same way as described above) to its recording, we have the reference microphone's signal calibrated from volts to Pa; then, we calculated the scaler factors—i.e., the sensitivity calibration factors—that the miniature microphones' signals required to match the the calibrated signal from the reference microphone.

For the calibrated signals from the three microphones, 1/6 octave band SPL was calculated, and the differences among them were qualified by their arithmetic difference at every frequency

band. This follows the standard calculation method of the objective occlusion effect (e.g., Saint-Gaudens *et al.* 2022); specifically, the objective occlusion effect is defined as the difference in SPL between the IEM and the OEM.

4.3.2.2 Voice quality and F0 metrics analysis

We used R (Team, 2022) for this part of analysis. Data processing was done with dplyr (Wickham *et al.*, 2022) and data visualization was done with ggplot2 (Wickham, 2016). More specific packages used are mentioned below.

a. Descriptive statistics

The complete data were grouped by the participants' sex and the recording microphone. Mean, SD and median were calculated for every metric in every group.

b. Modeling of the effects of microphone on voice quality measures

From a preliminary examination of variable distributions, jitter, shimmer and the SD of F0 (F0 SD) appeared heavily right-skewed as can be seen in Figure 4.2. Thus, a natural-log transformation for these variables was performed.

To account for sentence and subject-dependent variability, we analyzed the effect of microphone type (IEM, OEM, REF) on the voice quality metrics using linear mixed-effects (LME) modeling; detailed model constructions can be found in Tables 4.3 and 4.4. Specifically, our mixed-effects model have random intercepts and random slopes for microphone type, accounting for variability both within participants and within sentences. This modeling approach captures individual baseline differences across participants and sentences (random intercepts) while also accounting for how microphone positioning may affect each participant and sentence differently (random slopes). We built the models with lme4 (Bates *et al.*, 2015), summarized the pairwise estimate comparison results with emmeans (Lenth, 2022), and calculated p-values with lmerTest (Kuznetsova, Brockhoff & Christensen, 2017). For an easier interpretation, the LME results of the log-transformed metrics are reported as the pairwise comparison's percent change in

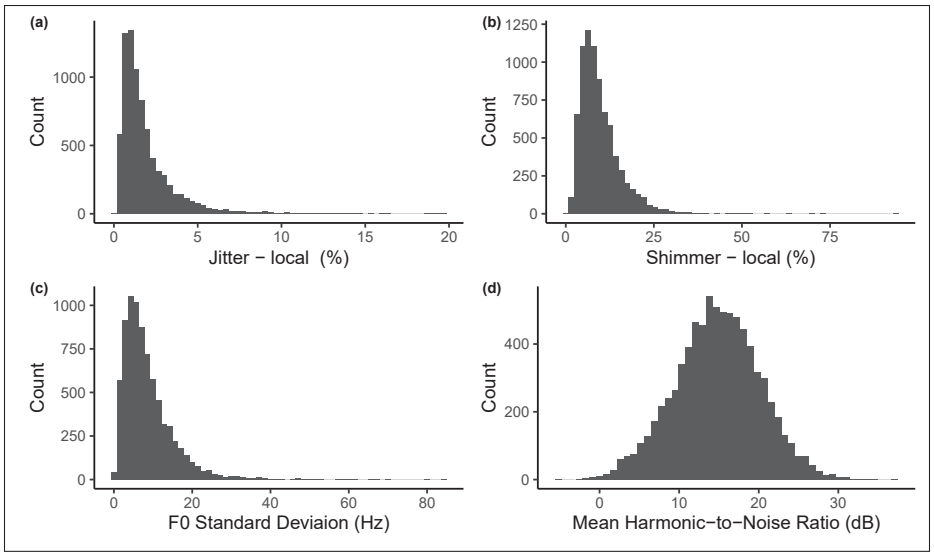


Figure 4.2 Histograms of four metrics from the complete dataset with the three microphones combined

The x-axis is the unit of the metric, and the y-axis is the count. Jitter, shimmer and F0 SD are right-skewed, in contrast to mean HNR, which is not skewed.

the original scale. A series of possible models were constructed for each metric with random intercepts and microphone type as slopes for participants and sentences. Since F0 is necessarily influenced by sex, all the possible models have sex as a predictor.

Table 4.3 LME model constructions for jitter, shimmer and HNR

Model	**Formula**
Model 0	Only random effects (REs): `y ~ 1 + (1 + mic | participant) + (1 + mic | sentence)`
Model 1	Fixed effect (FE) of microphone type and REs: `y ~ mic + (1 + mic | participant) + (1 + mic | sentence)`
Model 2	FEs of microphone type and sex, plus REs: `y ~ mic + sex + (1 + mic | participant) + (1 + mic | sentence)`
Model 3	FEs of microphone type and sex, as well as their interaction, plus REs: `y ~ mic*sex + (1 + mic | participant) + (1 + mic | sentence)`

Table 4.4 LME model constructions for F0

Model	Formula
Model 1	FE of sex and REs: `y ~ sex + (1 + mic | participant) + (1 + mic | sentence)`
Model 2	FEs of microphone type and sex, plus REs: `y ~ mic + sex + (1 + mic | participant) + (1 + mic | sentence)`
Model 3	FEs of microphone type and sex, as well as their interaction, plus REs: `y ~ mic*sex + (1 + mic | participant) + (1 + mic | sentence)`

We then performed LRTs as the model comparison method to select the best-fit model. Note that if the model without the microphone type as a fixed effect is selected (i.e., the first model in each series), it means that the microphone type does not have a statistically significant effect on the voice quality variable.

4.4 Results

4.4.1 Overall frequency-content comparison of the three microphones

In Figure 4.3, 1/6 octave-band transfer functions are plotted to show the frequency-dependent differences between different recordings. An octave band spans frequencies that double from bottom to top (such as 100-200 Hz or 500-1000 Hz), while 1/6 octave bands provide finer resolution by dividing each octave into six smaller frequency ranges. The 1/6 octave-band transfer function between the IEM and the REF is plotted in blue to show the difference in the frequency content between these two microphones when recording one's speech in the SpEAR database. The positive magnitude in the lower frequencies demonstrates the amplification from ear occlusion, and the overall downward trend and the negative magnitude above 2 kHz demonstrates the effect of the low-pass filtering. In Figure 4.3, the transfer function between OEM and REF is plotted in black. Their frequency contents are similar in the lower frequencies with OEM being slightly higher in magnitude, and OEM has a slightly lower magnitude in higher

frequencies. This is consistent with the red curve in Figure 4.3, which shows the difference between IEM and OEM.

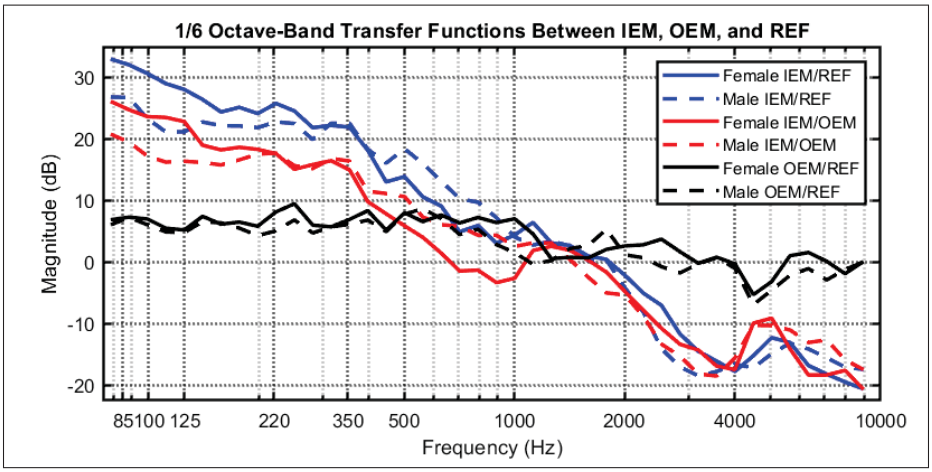


Figure 4.3 Spectral content differences between IEM, OEM and REF

X-axis shows frequency in log10-scale, y-axis shows difference in magnitude. Additional ticks are shown on the x-axis to facilitate result description. Colour indicates different microphone comparison; line type indicates sex.

While the overall trend remains the same, differences between males and females were observed under 300 Hz for the IEM versus REF comparison and under 200 Hz for the IEM versus OEM comparison. While the difference between IEM and the other two microphones continuously goes up with frequency decreases in females, in males they seem to be rather flat. Note that the F0 range of the female participants in this study was 125 ~ 350 Hz, corresponding to 23 ~ 15 dB of objective occlusion effect, and the males' was 85 ~ 220 Hz corresponding to 20 ~ 17 dB of objective occlusion effect. In general, larger magnitude of difference from occlusion occurred in the female F0 range than in the male F0 range. Additionally, females had an overall bigger objective occlusion effect as the maximum is 26 dB for females and 21 dB for males.

4.4.2 Descriptive statistics

The descriptive statistics for voice quality and F0 metrics, presented in Table 4.5, revealed some notable patterns. Overall, jitter metrics, shimmer metrics, and F0 SD showed a right-skewed distribution, indicated by the mean being larger than the median. Standard deviation for these metrics was also large in comparison to the corresponding mean, especially for the IEM condition.

Table 4.5 Descriptive statistics of voice quality and F0 measures from different microphones, grouped by sex

		Female									Male								
		IEM			OEM			REF			IEM			OEM			REF		
Metric		Mean	SD	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
Jitter	Local (%)	1.80	1.72	1.26	1.63	1.64	1.14	1.58	1.55	1.13	2.36	1.93	1.78	2.28	1.89	1.70	2.34	1.94	1.74
	Local (μs)	89.6	96.1	59.8	81.2	91.5	53.5	78.5	86.8	53.2	185	153	138	179	150	135	183	153	137
	RAP (%)	0.688	0.878	0.416	0.616	0.85	0.363	0.624	0.832	0.366	0.785	1.02	0.474	0.755	0.978	0.449	0.782	0.977	0.468
	PPQ5 (%)	0.702	0.734	0.469	0.635	0.758	0.403	0.636	0.800	0.398	0.802	0.744	0.583	0.792	0.764	0.572	0.815	0.774	0.586
Shimmer	Local (%)	9.48	5.98	8.17	7.46	4.71	6.37	8.87	4.52	7.83	14.1	7.82	12.6	9.50	4.84	8.68	11.2	5.27	10.3
	Local (dB)	0.831	0.483	0.732	0.697	0.408	0.594	0.823	0.410	0.739	1.22	0.615	1.12	0.866	0.434	0.807	1.01	0.470	0.950
	APQ3 (%)	3.20	2.31	2.54	2.42	1.97	1.88	2.75	1.90	2.23	5.81	3.77	4.95	3.66	2.55	3.03	4.17	2.69	3.52
	APQ5 (%)	5.10	3.93	4.07	3.65	2.51	2.99	4.61	3.04	3.77	8.32	5.53	7.00	5.02	2.93	4.44	6.19	3.72	5.40
	APQ11 (%)	8.38	6.23	6.51	6.66	4.65	5.46	9.13	5.92	7.51	10.4	7.36	8.75	8.47	5.07	7.44	10.8	6.76	9.44
HNR	Mean (dB)	18.1	5.41	18.3	17.4	4.54	18.0	14.8	4.08	15.3	10.3	4.83	10.5	12.8	3.83	13.1	10.9	3.45	11.0
F0	Mean (Hz)	213	33.4	211	214	33.6	211	214	33.6	212	132	22.6	128	132	22.7	128	132	22.8	128
	SD (Hz)	10.3	7.64	8.7	9.58	7.42	7.80	9.19	7.42	7.45	6.73	5.03	5.66	7.01	5.22	5.83	6.86	5.29	5.66
	Min (Hz)	196	32.9	195	198	33.1	198	199	33.2	199	121	22.5	116	121	22.8	115	121	22.7	115
	Max (Hz)	234	36.3	233	232	35.6	229	232	36.1	229	142	25.3	139	143	25.2	141	143	25.4	141

For jitter, male participants showed similar values across different microphone conditions, with minimal variation across metrics. In contrast, female participants exhibited higher jitter values in the IEM condition compared to the REF and the OEM. The OEM and the REF conditions were similar, though OEM showed slightly higher values in the two local metrics. When it comes to shimmer, for both males and females, the values were generally comparable between the IEM and the REF, with the IEM showing slightly higher values overall, while the OEM consistently recorded the lowest values. The SD for shimmer was also larger in the IEM condition compared to the REF and the OEM.

F0 metrics were generally consistent across microphone conditions, though slight differences were observed in female participants, with the IEM showing more variation compared to the OEM and the REF. For HNR, the IEM and the OEM showed higher values than the REF among female participants. In male participants, the OEM recorded the highest HNR values, with the

IEM and the REF being similar. The SD was again larger in the IEM condition compared to the OEM and the REF.

4.4.3 LME modeling

Tables 4.6 and 4.7 show the results of the LRTs for each metric. The results revealed a significant main effect of microphone condition on voice quality measures for all metrics except mean and maximum F0, with p-values ranging from <0.0001 to 0.049 when comparing models with and without microphone condition as a predictor. A main effect of sex was also evident for jitter, shimmer, and HNR, with p-values ranging from 0.003 to 0.274 when comparing models with and without sex as a predictor. Additionally, an interaction effect was observed for jitter, HNR, F0 SD, and weakly for minimum and maximum F0, with p-values ranging from 0.014 to 0.208; the interaction effect is not observed in shimmer metrics.

Table 4.6 LRT results for different LME models for jitter, shimmer and HNR

P-values indicate significance of fit improvement by adding terms.

	Jitter				**Shimmer**					**HNR**
Pr(< Chisq)	Local	Local (abs)	RAP	PPQ5	Local	Local (abs)	APQ3	APQ5	APQ11	Mean
Model 0: `y ~ 1`										
Model 1: `y ~ mic`	0.016	0.017	0.012	0.041	<0.0001	<0.0001	<0.0001	<0.0001	0.00013	<0.0001
Model 2: `y ~ mic + sex`	0.068	0.007	0.062	0.274	0.0097	0.011	0.003	0.0096	0.0073	0.0127
Model 3: `y ~ mic*sex`	0.014	0.017	0.15	0.128	0.421	0.395	0.376	0.417	0.296	0.021

Table 4.7 LRT results for different LME models for F0

P-values indicate significance of fit improvement by adding terms.

	F0			
Pr(< Chisq)	Mean	Min	Max	SD
Model 1: `y ~ sex`				
Model 2: `y ~ mic + sex`	0.88	0.06	0.59	0.049
Model 3: `y ~ mic*sex`	0.99	0.21	0.14	0.014

Below, we show the detailed results from the best-fit LME models in forest plots. We compared the difference between each microphone and sex indicated on the y-axis; the unit is indicated on the x-axis. Within each figure, each color represents one metric, each dot represents the estimated difference from the comparison, the lines to the side of each dot represent the 95% CI, and p-values are shown above each line.

4.4.3.1 Jitter

Figure 4.4 shows that jitter metrics were similarly affected by microphone condition when grouped by sex, with variations observed within the 95% CI. For females, an increase in jitter was noted in the IEM condition compared to OEM (12.4% to 17.8%) and REF (14.1% to 17.1%), with no significant difference between OEM and REF. For males, no statistically significant differences across the microphones were observed. Additionally, males generally exhibited higher jitter values than females, particularly in local (absolute) metrics.

4.4.3.2 Shimmer

Figure 4.5 shows significant differences across different microphones and between sexes for shimmer metrics . Notably, the IEM shows higher values than OEM, with statistically significant increases of 25.7 ~ 44.2% across all metrics. The OEM shows lower values than the REF, with statistically significant decreases of -36.3 ~ -18.5% across all metrics. When comparing the IEM to the REF, the differences are less pronounced and inconsistent,in the –10.1 ~ 20.1% range, not statistically significant except for APQ3 ($p = .0116$). The main effect of sex is also evident, with statistically significant higher shimmer seen in males across all metrics; the increase ranges from 25 to 50.7%.

4.4.3.3 HNR

Figure 4.6 shows that for mean HNR, while the increase seen in the OEM compared to the REF is similar in both sexes (F: 2.63 dB; M: 1.94 dB), difference is seen in the IEM compared to the

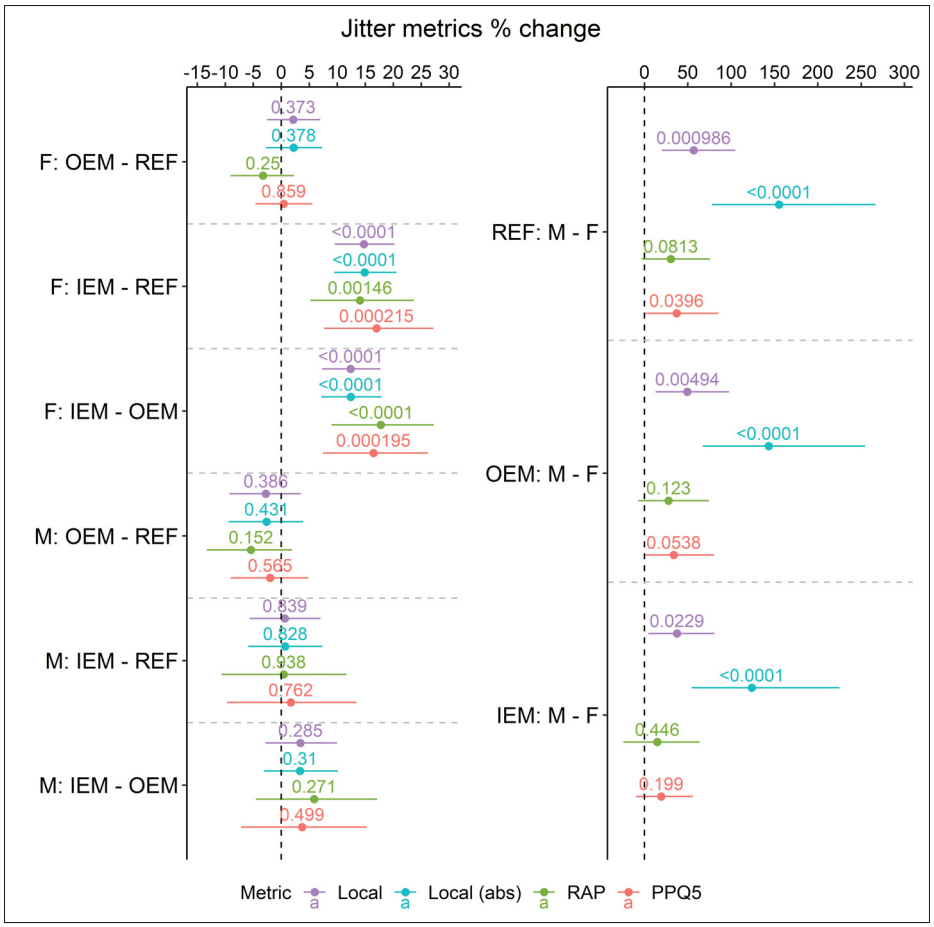


Figure 4.4 Jitter metrics percent change estimated by Model 3

The left side of the figure compares the effects of microphone condition within each sex, the right side of the figure compares the effects of sex within each microphone condition. For example, F: IEM – OEM reads as the difference between IEM and OEM for female participants.

REF and the OEM. For females, the IEM is comparable to the OEM but is higher than the REF by 3.29 dB ($p < .001$). For males, in contrast, the IEM is comparable to the REF but lower than the OEM by 2.53 dB ($p < .05$). Overall, females have statistically significantly higher HNR than males in all three microphones, ranging from 3.91 ~ 7.79 dB; the difference is largest in the IEM.

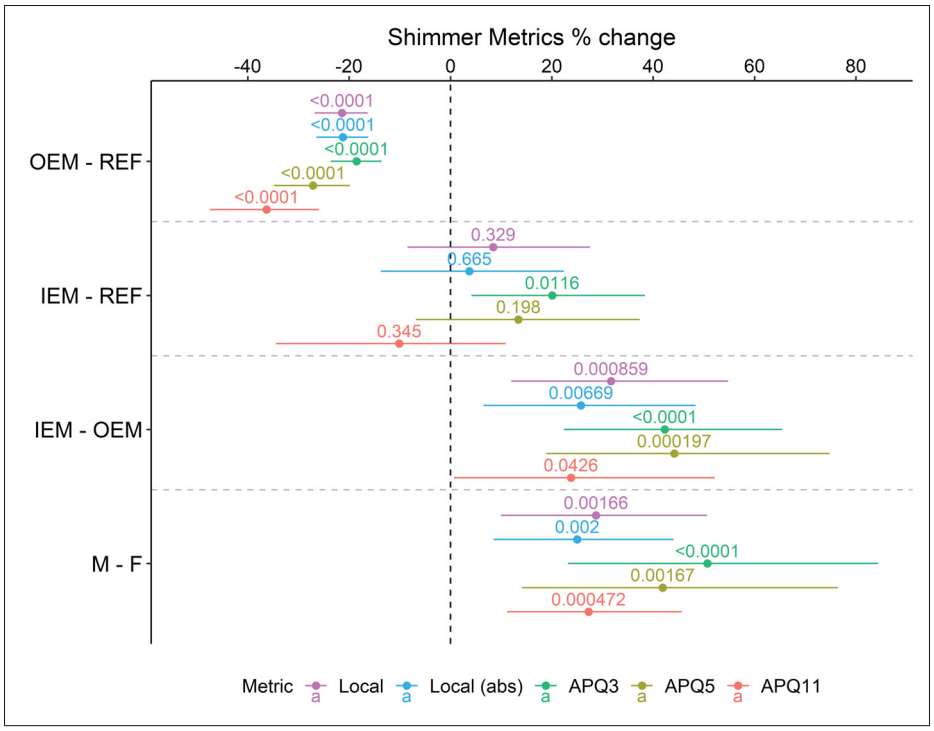


Figure 4.5 Shimmer metrics percent change estimated by Model 2

For example, IEM – OEM reads as the difference between IEM and OEM.

4.4.3.4 F0

As shown in Figure 4.7, for F0 metrics, the effect of microphone is different in females versus males. For males, no significant effect of microphone is seen in any metric. For females, no effect of microphone is seen for F0 mean, but there appears to be small and weak effect for F0 max and min comparing the IEM to the REF and the OEM. The IEM has lower F0 min and higher F0 max compared to the REF and the OEM; the estimated differences are within 3 Hz on average. For F0 SD, the IEM and the OEM have higher values than the REF by 15.6% ($p < .0001$) and 5.68% ($p < .005$) respectively. A main effect of sex is also observed across all metrics, with females having higher values than males.

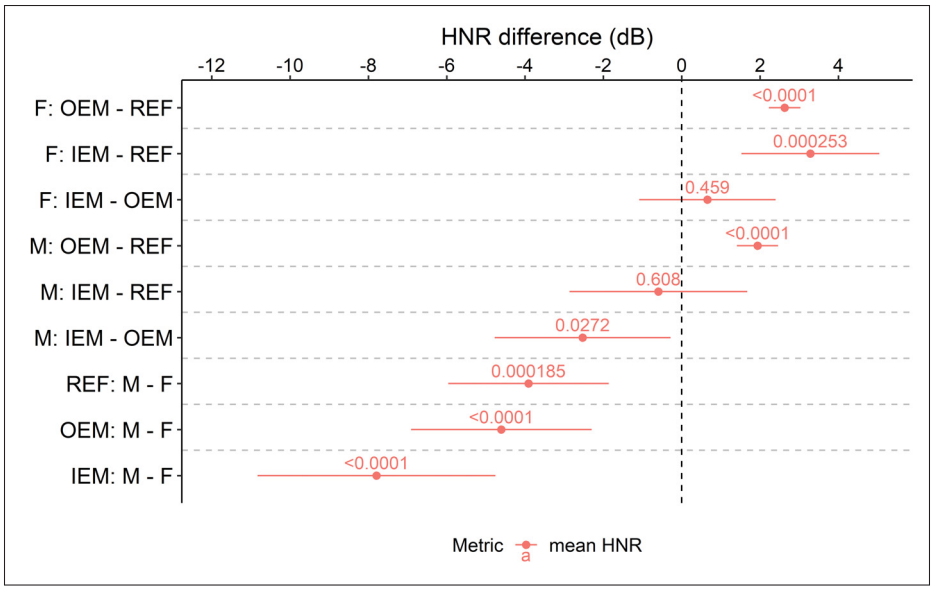


Figure 4.6 HNR differences estimated by Model 3

For example, IEM: M - F reads as the difference between males and females in IEM.

4.5 Discussion

To summarize, microphone placement affects voice quality and F0 measurements, as demonstrated through both descriptive statistics and LME modeling. Specifically, higher and more variable jitter values were recorded with the IEM compared to the OEM and the REF in females, with no significant difference between the OEM and the REF jitter. In females, the IEM also recorded a lower minimum F0, a higher maximum F0, and an increased SD compared to the OEM and the REF. No such effects were observed in males. These results suggest that placing the microphone within an occluded ear canal introduces variability or artifacts that impact the measurements of jitter and F0, with this effect being sex-dependent. Furthermore, both the REF and the IEM recorded similar shimmer values, which were greater than those for the OEM. Finally, the OEM recorded higher HNR values compared to the REF. However, in females, the IEM recorded higher HNR than the REF and was comparable to the OEM, while in males, the IEM recorded HNR values similar to the REF but lower than the OEM. This suggests that placing

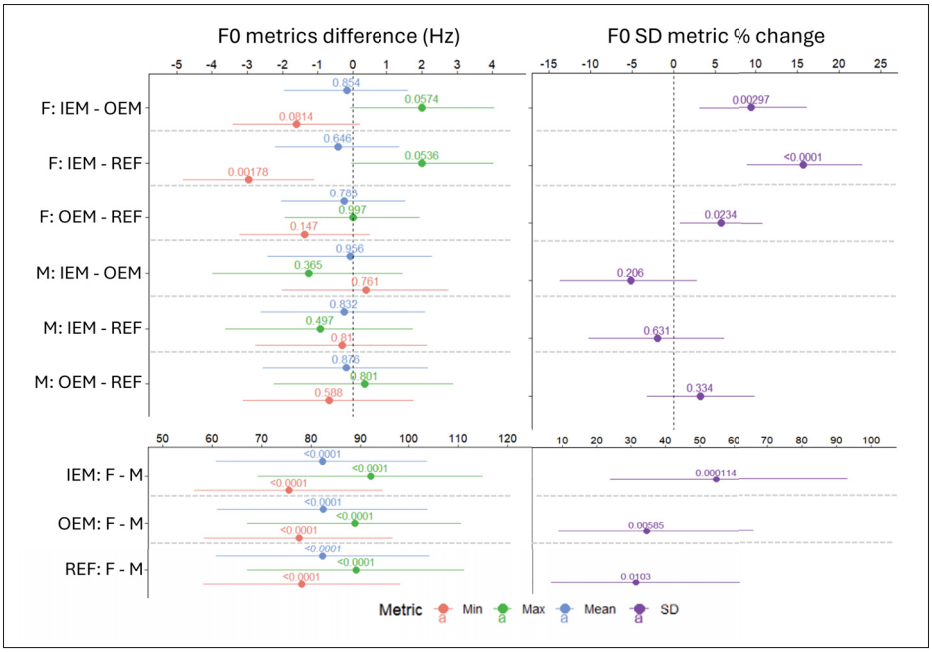


Figure 4.7 F0 metrics differences estimated by Model 3

The left side plots F0 mean, max, and min in absolute difference (Hz); the right-side plots F0 SD in percent change. The upper section plots the differences between microphones given a sex; the lower section plots the differences between sexes given a microphone. For example, IEM: F - M reads as the difference between females and males in the IEM. Note that the sex comparison direction is switched here for easier interpretation in positive numbers for F0 SD percent change.

the microphone outside of the ear canal reduces shimmer and increases HNR. A sex-dependent effect is observed for HNR, with the IEM recording higher HNR in females but not in males.

The observed differences in HNR may depend on how microphone placement influences the recording of different frequency ranges. Specifically, the distinction between OEM and REF conditions is likely due to the head-related transfer function (HRTF) between the mouth and ear. Dunn & Farnsworth (1939) showed that compared to a frontal microphone placed at 30 cm from the mouth, a microphone placed at 135 degrees azimuth and 15 cm from the mouth amplifies frequencies under 1 kHz by 3~5 dB and attenuates frequencies above 1 kHz by 1~8 dB. This location is comparable to the OEM location, and a similar attenuation is observed in its

transfer function (see in Figure 4.3). This attenuation of high frequencies due to shifted sound orientation has been observed in more recent research (Kato, Takemoto, Nishimura & Mokhtari, 2010; Leishman, Bellows, Pincock & Whiting, 2021; Bellows & Leishman, 2022). This occurs because high-frequency energy is more directionally focused toward the front of the source, while the low-frequency energy radiates omnidirectional around the source (Moriarty, Ananthanarayana & Monson, 2024). When comparing OEM and REF, we saw an amplification in the lower frequencies in OEM due to the proximity to the source; together with the attenuation of higher frequencies, these two factors could increase HNR, as lower frequencies have more energy on the harmonics and higher frequencies have more noise.

Following this reasoning, however, one may see an apparent discrepancy when explaining the effect of IEM on HNR. Like OEM, IEM also has increased levels in lower frequencies and decreased levels in higher frequencies; in fact, the IEM's amplification and attenuation effects are more than 15dB stronger than the OEM. Yet, unlike the effect of the OEM, the effect of the IEM is sex dependent, with increased HNR seen in females and the opposite in males. This interaction effect could be further explained by the difference between the OEM and the IEM low-frequency amplification and the male speakers' voice characteristics. Shown in Figure 4.3, the low-frequency amplification is more uniform in the OEM than in the IEM. In the IEM, the lower the frequency, the higher the amplification. In this case, if there is a source of anharmonic sound that is lower than F0 (also the first harmonic), it would be reasonable if the HNR decreases significantly. As shown in Table 4.5, HNR of male speakers in this dataset is on average 4 dB lower than females. This difference aligns with informal observations during data exploration that suggested more instances of vocal fry in male recordings, which produces an anharmonic sound lower than F0, ranging from 2 to 78 Hz (Hollien, 1974); however, this preliminary observation would require systematic perceptual evaluation to confirm.

In summary, while the OEM records higher HNR, likely due to filtering of high frequencies and amplification of lower ones, the IEM's effects are more nuanced and appear to differ based on sex. Further investigation into the specific characteristics of the noise components and

their interaction with microphone placement could provide deeper insights into these observed patterns.

The difference seen in shimmer may be explained similarly to HNR, as shimmer is also calculated from the amplitude of the sound signals. Like HNR, shimmer was significantly different between the OEM and the REF recordings. This difference could stem from the HRTF, where attenuation occurs for frequencies above 1 kHz, or it might be due to reduced sensitivity in the OEM compared to the REF. Additionally, the REF microphone, being farther from the sound source, may have been more influenced by sound reflections from the booth walls, as the mouth-to-microphone distance affects how reverberant noise affects the signal. In contrast, the OEM microphone, positioned closer to the mouth, would be less affected by the reverberant field. However, since the recording environment had a low reverberation level, we do not expect that distance played an important role here. Regardless of the true cause, what the results indicate is that the increase in shimmer that we see in the IEM compared to the OEM is not because that IEM captured significantly more shimmer, but because OEM captured significantly less shimmer. Indeed, the IEM does not capture more shimmer than the REF on average; although it did show higher variability than the REF. Given these factors, the nature of shimmer in the IEM may differ from that in the REF and OEM conditions, warranting further investigation.

The limited effect of the IEM on shimmer compared to the REF is in itself an intriguing observation. High frequency contents are more attenuated in the IEM than in the OEM, and the IEM and the OEM have similar sensitivity. However, shimmer did not decrease in the IEM compared to the REF. One possible explanation to this apparent discrepancy is that the IEM amplifies lower frequencies more than the OEM, due to proximity to the sound source. Thus, it is possible that IEM's overall amplification translates to shimmer amplification, making the peaks in each period more prominent (see example in Figure 4.8).

The results for jitter and F0 are similar in that differences are observed between the IEM and the other two microphones but only in female participants. The increase in jitter in the IEM suggests higher variability in the period lengths, which also manifests as greater variability in F0. Indeed,

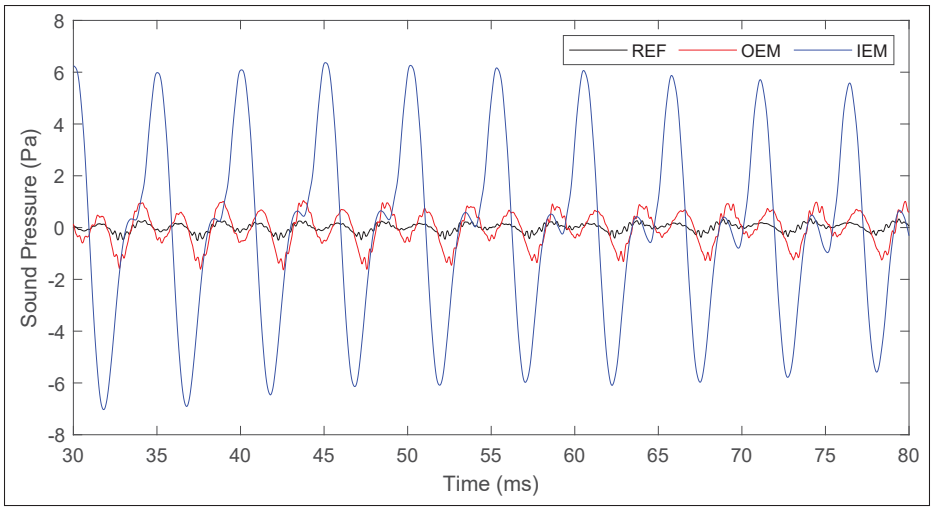


Figure 4.8 The waveforms of /i/ uttered by a female participant

The x-axis is time (ms), and the y-axis is sound pressure. Fewer high frequency components are visible in IEM than in the two other microphones, and the peaks are more prominent with less variations. OEM also has fewer higher frequency components but in a lesser degree.

in F0 metrics, higher maximum, lower minimum, and larger SD are observed with the IEM. It is unclear why there is a sex effect; future research may investigate whether this effect is dependent on one's pitch range.

Additionally, some notable secondary findings were observed. Jitter and shimmer are generally higher than typically seen in populations without voice disorders when measured using sustained vowels. For example, the pathological boundary for jitter is 1.04% (local) and 83.200 μs (local, absolute), and 3.810% (local) and 0.350 dB (local, absolute) for shimmer with the Multidimensional Voice Program(Zelcer, Henri, Tewfik & Mazer, 2002), yet the values in this study are higher. This is expected, as there is more variation in continuous speech than in sustained vowels. Significant sex-based differences were also noted: males exhibited higher jitter and shimmer and a lower HNR compared to females. This difference is also observed in a study done with 49 speakers (Brown & Sonderegger, 2024). These differences were particularly pronounced in the IEM condition for F0 SD and HNR.

The objective occlusion effect, measured from the spectral difference between the OEM and the IEM, aligns with reports in the literature. A sex-based difference was also noted: females exhibited a higher objective occlusion effect than males below 350 Hz and a lower objective occlusion effect between 350 and 2000 Hz than males. Further research is needed to explore the potential effects of ear occlusion and its interaction with sex on voice quality and F0.

However, due to the relatively small sample size, we cannot definitively attribute the observed differences to sex alone; they may also result from individual variations in the fit of the equipment or the degree of OE. Future studies should aim to collect data with more participants with similar proportion of males and females, and explore the effect of ear occlusion. While overall objective measures of ear occlusion across frequencies and F0 ranges by sex were reported in our study, directly modeling these complex effects they may have on voice quality measurements in addition to microphone placement requires careful consideration. The occlusion effect varies across frequency ranges, making it challenging to determine an appropriate single metric for statistical modeling. We envision a single-value indicator based on the F0 range of the speaker or the utterance instance; however, this requires further validation with future studies.

4.6 Conclusion

This study demonstrates that microphone placement significantly affects voice quality and F0 metrics with notable differences between sexes. Microphones placed within an occluded ear canal introduce variability and artifacts, particularly in female participants, while microphones placed outside the ear canal yield lower shimmer and higher HNR values. The effects of the outside microphone placement can be explained by the head-related transfer function between its location and that of the reference microphone. The effects of the IEM placement could be explained by the characteristics of bone-and-tissue conducted speech and the effect of ear occlusion. The observed sex-dependent effects, particularly for jitter, HNR, and F0 variability, suggest that microphone placement interacts differently with male and female vocal characteristics, potentially due to physiological differences such as F0 range and earpiece fit. However, given our limited sample size and unbalanced sex distribution, these sex-based findings

should be interpreted as preliminary observations that require further investigation with larger, more representative samples. Our findings suggest that standards that have been established with transitional in-front-of-the-mouth microphones may not be directly applied to data collected with hearables. Understanding these differences, identifying their causes, and exploring whether they can be mitigated are crucial steps for effectively using hearables as a tool for health monitoring. This study lays the groundwork for future research on refining voice quality measurement in hearables.

4.7 Author Declarations

The authors have no conflicts of interest to disclose. The data of this study are available upon request. This research project received ethics approval from École de technologie supérieure (H20170103).

4.8 Acknowledgments

This research was supported by Natural Sciences and Engineering Research Council of Canada (NSERC), CIRMMT, and the École de technologie supérieure's Marcelle Gauvreau Engineering Research Chair in Multimodal Health Monitoring and Early Disease Detection with Hearables. We thank Daniel Zhou and Philippe Rochefort for annotating the data.

CONCLUSION AND RECOMMENDATIONS

This thesis investigated how speech production is shaped by a combination of environmental, physiological, and technological factors, with a particular focus on conditions that reflect communication in real life. By examining speech across diverse sensory conditions—from immersive virtual environments to noisy settings with ear occlusion and hearing impairment—this work challenges traditional assumptions about speech motor control and demonstrates that variability across individuals and contexts is not merely noise in the data, but rather a fundamental feature of how humans adapt their vocal behavior. Across the three studies, the work demonstrated that speech control is not solely an auditory process but is shaped by multisensory input and individual differences, while also underscoring the methodological impact of measurement technology on speech analysis. Together, these findings call for a shift away from group-level, linear models toward individualized, dynamic frameworks that account for the complex interactions between sensory feedback, speaker characteristics, and the tools we use to measure speech.

5.1 Multisensory integration and temporal dynamics in speech control

The first study showed that while auditory information remains the dominant cue in regulating speech level, visual context can also shape vocal behavior, particularly in the early stages of speech production. Importantly, the temporal pattern revealed that visual information exerted its influence early and then diminished, while auditory effects emerged more gradually and persisted throughout speech production. This suggests that different sensory modalities may operate on different timescales in speech motor control—a finding with important implications for models of sensorimotor integration. These findings extend theoretical models of multisensory integration in speech motor control and provide practical insights for the design of immersive virtual environments, where careful manipulation of auditory and visual cues could be leveraged

to support more natural vocal behavior in virtual communication, training, and therapeutic contexts.

5.2 Individual variability and nonlinear effects in auditory feedback control

The second study revealed that speech production under noise, ear occlusion, and hearing impairment is far more variable and non-linear than previously assumed. Rather than simple additive or linear relationships, the results demonstrated substantial individual differences in how speakers respond to ear occlusion, with some increasing and others decreasing their speech levels. Furthermore, the preliminary investigation into hearing impairment uncovered a categorical shift around PTA = 15 dBHL, suggesting qualitatively different control strategies above and below this threshold. These nonlinearities may help reconcile conflicting findings in the literature, which have typically assumed linear models and overlooked individual variability. By introducing HIBiSCus, the work not only addressed long-standing gaps in the literature on speech production with ear occlusion but also provided a rich, openly available resource for future research on hearing, feedback control, and speech in French and English. The corpus enables investigations into how bilingual speakers adapt their vocal behavior across languages and sensory conditions—a question that remains largely unexplored. The results highlighted the importance of individualized modeling approaches and pointed toward potential new applications in early detection of hearing impairment through speech production patterns and improvement of in-ear device fitting procedures. If longitudinal speech patterns could serve as sensitive indicators of auditory changes, hearable devices might enable continuous, ecologically valid monitoring of hearing health in daily life—complementing traditional clinical assessments.

5.3 The role of measurement technology

The third study demonstrated that microphone placement inside and around the ear, as used in hearables, has a significant influence on acoustic measurements of voice quality and F0.

Microphones placed within the occluded ear canal introduced greater variability and artifacts, particularly in female participants, while external placements yielded more stable measurements. These sex-dependent effects may reflect physiological differences in F0 range and earpiece fit, though the preliminary nature of these findings warrants further investigation with larger samples. The results highlight the need for methodological refinements and new standards when using hearables for continuous voice evaluation. Existing standards established with traditional in-front-of-the-mouth microphones cannot be directly applied to data collected from hearables, as the acoustic properties of bone-and-tissue conducted speech and the effects of ear occlusion fundamentally alter the signal. Beyond methodological implications, the findings underscore both the promise and the challenges of hearable technology for non-invasive health monitoring. If these measurement artifacts can be systematically characterized and compensated for, hearables could provide unprecedented access to naturalistic speech data in everyday contexts—enabling longitudinal tracking of vocal health, cognitive state, and auditory function.

5.4 Individual differences, dynamic control, and ecological validity

Taken together, these studies contribute to a more nuanced understanding of speech production in real-world and technologically mediated contexts. Several overarching themes emerge: **Individual variability as a core feature.** Across all three studies, substantial individual differences were observed—whether in responses to visual immersion, reactions to ear occlusion, or sex-dependent effects of microphone placement. Rather than treating this variability as experimental noise to be averaged away, this work demonstrates that individual patterns are informative and theoretically meaningful. They likely reflect differences in sensorimotor control strategies, auditory sensitivity, physiological characteristics, and prior experience.

Nonlinearity and context-dependence. The categorical shift in hearing impairment effects, the temporal dynamics of audio-visual integration, and the interaction between occlusion and hearing loss all point to nonlinear, context-dependent processes. Speech motor control cannot

be adequately captured by simple linear models; instead, frameworks must account for threshold effects, temporal dynamics, and interactions between multiple factors.

The inseparability of measurement and phenomenon. The third study serves as a crucial reminder that our understanding of speech is inextricably linked to how we measure it. As hearable technologies become more prevalent in speech research and clinical applications, establishing robust methodological standards and understanding the artifacts introduced by these devices is essential.

Bridging theory and application. Each study demonstrates how basic research on speech motor control can inform practical applications—from designing better virtual environments to detecting hearing impairment earlier to developing standards for health monitoring technologies. This bidirectional relationship between theory and practice strengthens both.

5.5 Future directions

Looking forward, several promising directions emerge from this work:

Temporal dynamics of multisensory integration. The finding that visual and auditory information operate on different timescales invites further investigation. How do these temporal patterns vary across tasks, speaking styles, and individuals? Can computational models of sensorimotor control account for these dynamics?

Expanded corpus-based approaches. While HIBiSCus provides valuable data on speech level, future work should examine a broader range of acoustic and articulatory features—including segmental characteristics, prosody, voice quality, and articulatory kinematics—to build a more complete picture of how speech adapts to sensory perturbations. Expanding the corpus to include more diverse populations (age groups, clinical populations, language backgrounds) would also enhance its utility.

Individualized models and prediction. Can statistical approaches leverage individual baseline

patterns to predict responses to new conditions or detect early signs of hearing decline? Developing personalized models of speech motor control could enable tailored interventions in clinical and assistive technology contexts.

Standards for hearable-based evaluation. Establishing robust standards for voice quality and F0 measurement from hearables will require systematic investigation of how different device designs, microphone placements, and signal processing approaches affect acoustic measurements. Ideally, compensation algorithms could be developed to map hearable-recorded speech onto standardized reference spaces.

Real-world deployment and validation. Ultimately, the value of hearable-based monitoring depends on its performance in naturalistic settings over extended periods. Longitudinal studies that track speech patterns alongside clinical measures of hearing, vocal health, and cognitive function will be essential for validating these technologies and translating research findings into practice.

5.6 Concluding remarks

This thesis demonstrates that speech production is shaped by a dynamic interplay of sensory inputs, individual characteristics, and measurement methods. By moving beyond traditional laboratory paradigms and embracing the complexity of real-world communication, this work contributes to both theoretical models of speech motor control and practical applications in virtual communication, hearing health, and hearable technology. The findings underscore that to truly understand how humans speak, we must account for the full richness of the contexts in which speech occurs—and the tools through which we observe it.

APPENDIX I

SUPPLEMENTARY MATERIAL FOR CHAPTER 1

1. Full generalized additive mixed-effects model summary

```
Family: gaussian
Link function: identity
Formula:
spl_f.c ~ s(reftime, k = 5) +
s(reftime, aud, k = 5, bs = "sz") +
s(reftime, vis, k = 5, bs = "sz") +
s(reftime, vis, aud, k = 5, bs = "sz") +
s(reftime, Participant, bs = "fs", m = 1, k = 5)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.008089   0.006759   1.197    0.231

Approximate significance of smooth terms:
                           edf Ref.df      F p-value
s(reftime)               3.840  3.862  154.6  <2e-16 ***
s(reftime,aud)           9.767  9.974 2837.3  <2e-16 ***
s(reftime,vis)           9.887  9.995  213.8  <2e-16 ***
s(reftime,vis,aud)      16.460 16.935  105.9  <2e-16 ***
s(reftime,Participant) 89.764 99.000  186.9  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
R-sq.(adj) =  0.236   Deviance explained = 23.6%
fREML = 3.7853e+05  Scale est. = 3.0789    n = 190882
```

2. Generalized Additive Mixed Models Comparison

Model name	Model description	AIC	Stepwise AIC difference	fREML Score	Stepwise fREML difference	p-value	
1. Model	Full model	756499.0		378531.9		*Compared to Model 1*	
2. Model_no_AudVisInt	Removed Aud_Vis interaction	758260.2	1761.2	379376.0	844.1	< 0.001	*Compared to Model 2*
3. Model_no_Vis	Removed Vis main effect and Aud_Vis interaction	760383.9	2123.7	380412.1	1036.1	< 0.001	< 0.001
4. Model_no_Aud	Removed Aud main effect and Aud_Vis interaction	784416.4	26156.2	392420.6	13044.6	< 0.001	< 0.001
5. Model_no_AudVis	Removed Vis and Aud main effects and Aud_Vis interaction	786239.1	27978.9	393307.1	13931.1	< 0.001	< 0.001

Summary of the model comparison results.

The absolute AIC score and minimized smoothing parameter selection score (from fREML) are provided. The Chi-Square test uses the minimized smoothing parameter selection score with the compareML() function from the itsadug package. The stepwise differences are calculated as such: Model 2 was compared to Model 1, and Model 3, 4, 5 were compared to Model 2. This is because Model 2 removed the interaction term, and Model 3, 4, 5 built upon the removal of the interaction term.

3. Empirical curves

Here we include the empirical speech level curves in every Aud-Vis condition. As reported in Section 2.4.2, we show two versions of the visualization; one grouped by Aud and one grouped by Vis. Below, Figure I-1 is plotted with the sonorant, LOESS-treated data points, and Figure I-2 is plotted with the raw sonorant data points. We can see that the curves did not change in their shapes but only in their standard errors shown by the grey bands. This further confirms the validity of the LOESS smoothing treatment. Visual comparison also confirmed that the fitted smooth curves from our Generalized Additive Mixed Model closely aligned with these empirical trends, particularly in their overall shape and inflection points. This strong agreement provides additional validation for the GAMM's ability to robustly capture the underlying temporal patterns and interactions relevant to our research questions.

4. Interaction curves similarity

In Figure I-3, we replotted the interaction effect curves and annotated them for the similarity that we see in each column. The curves grouped by Vis (on the bottom) were rearranged in their sequence to match with the top row grouped by Aud. This sequence shows the movement of Vis-A and Vis-G across Aud conditions and the movement of Aud-C and Aud-G across Vis conditions as mentioned in Section 2.4.4. Comparing each of the two figures in the same column, we can see overall resemblances. The pairs of curves that are similar are annotated with the same symbol at the end; the two curves that are not annotated in each pair are the same curve. This information is further summarized in the right side Table (1), which is then transformed to Table (3) by incorporating the partial effects on speech levels from Aud or Vis shown in Table (2). In Table (2), the letter encoding for the conditions was matched with a numerical encoding, indicating their effects on speech levels; the larger the number, the higher the speech level.

To more formally assess the observed similarities, we computed the pairwise Euclidean distances between curves; the resulting heatmap is presented in Figure I-4. As shown, the three pairs of

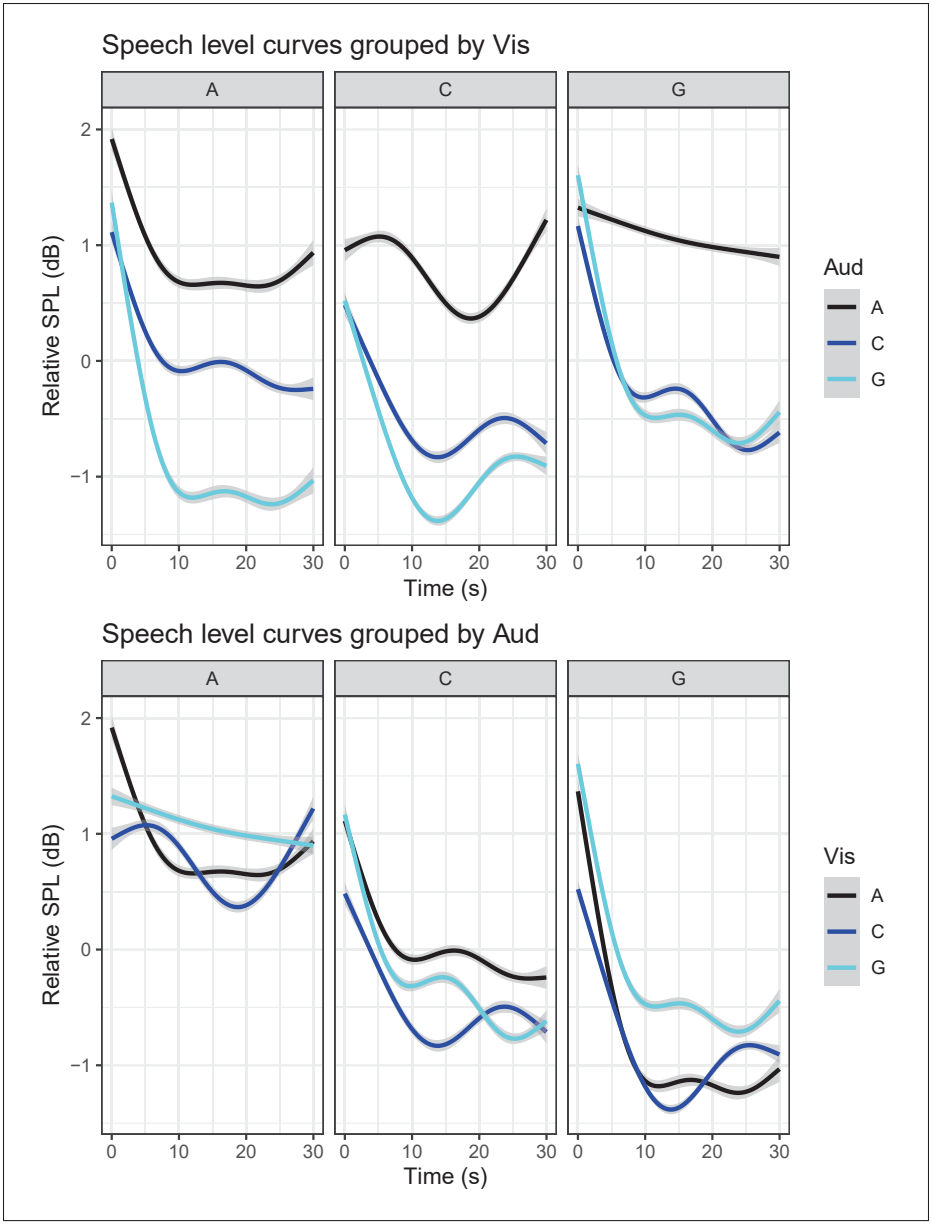


Figure-A I-1 Empirical speech level curves with LOESS-treated sonorant data points

curves highlighted in bright yellow have the smallest distances and are the same pairs presented in Figure I-3.

When participants spoke under the condition where Aud is one level higher than Vis, the interaction effects followed the circle trend; when the condition is that Aud is one level lower than

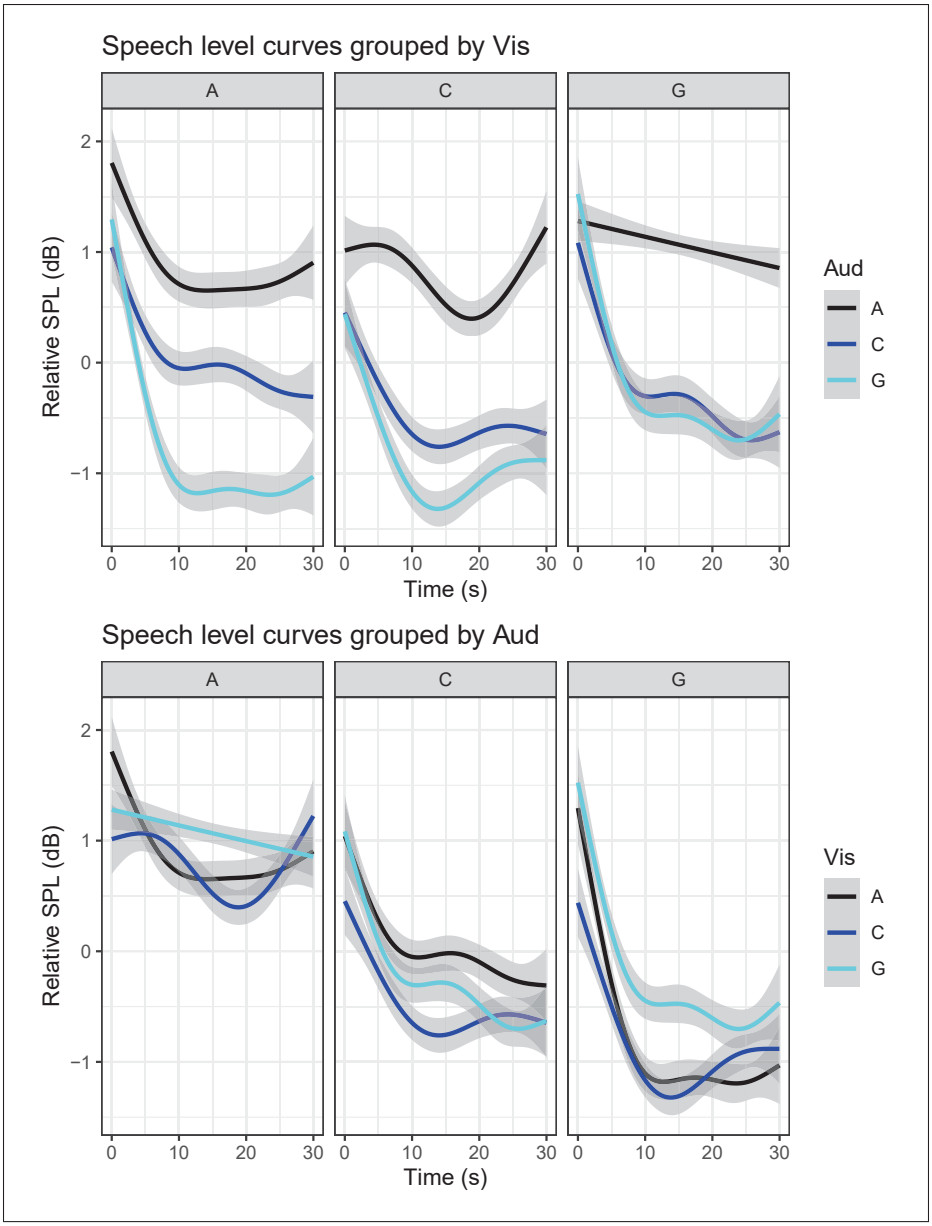
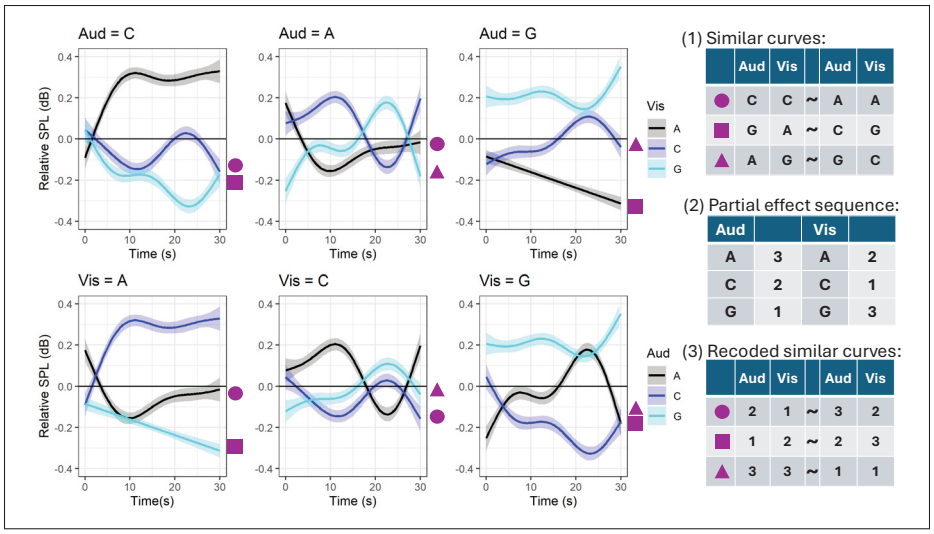


Figure-A I-2 Empirical speech level curves with raw sonorant data points

Vis, the interaction effects follow the square trend; and when the Aud and Vis levels matched, the interaction effects follow the triangle trend. These findings show that participants were integrating both Aud and Vis information and systematically adjusting to the combined effects of these two streams of information.



	Aud	Vis		Aud	Vis
●	C	C	~	A	A
■	G	A	~	C	G
▲	A	G	~	G	C

Aud		Vis	
A	3	A	2
C	2	C	1
G	1	G	3

	Aud	Vis		Aud	Vis
●	2	1	~	3	2
■	1	2	~	2	3
▲	3	3	~	1	1

Figure-A I-3 The similarities among the Aud-Vis interaction effects curves

On the left, we included the six sub-plots from Figure 2.8 and Figure 2.10. The symbols at the end of a curve mark out the pairs that are similar to each other. The Aud-Vis condition of these curves are listed on the right side table (1). The numerical sequence of speech levels in Aud and Vis partial effects is summarized in right side table (2). The re-coded Aud-Vis conditions for the similar curves are in right side table (3).

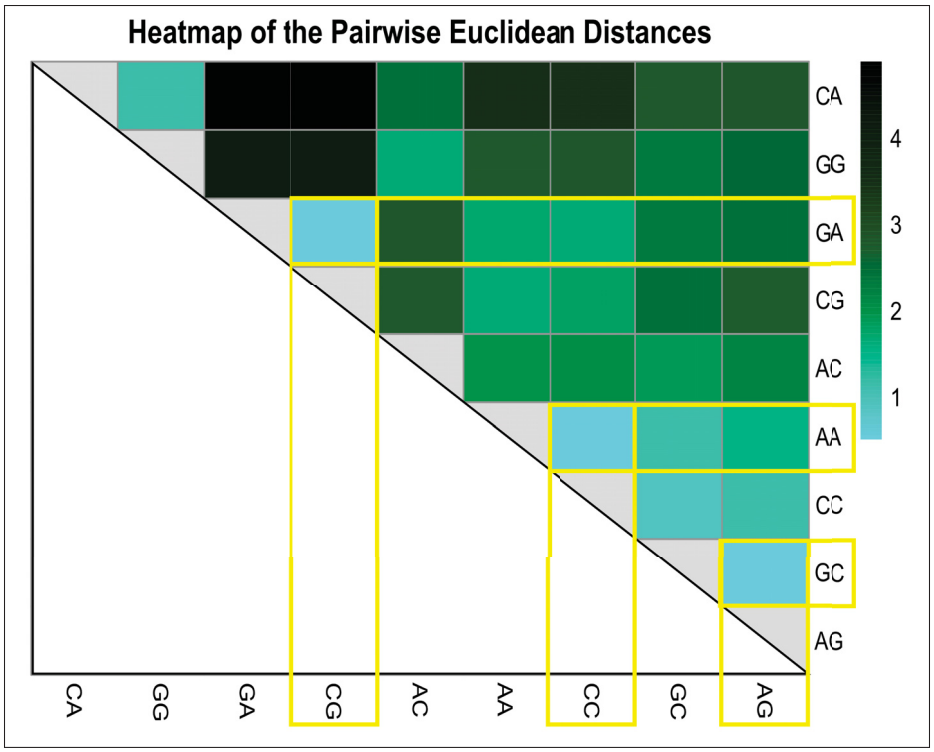


Figure-A I-4 Heatmap of the pairwise Euclidean distances between interaction curves

Color intensity reflects distance magnitude, with lighter (blue) shades indicating greater similarity. Curve labels follow the format "AudVis," where each letter represents the Aud and Vis conditions (e.g., CC = Aud-C Vis-C).

APPENDIX II

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

1. Descriptive statistics on speech level

Table-A II-1 Descriptive statistics with any-frequency grouping

Noise	Occlusion	Hearing Group	N	Mean (95% CI)	SD	Range
NN	OE	N	30	50.8 (49.4-52.3)	3.858276	44.6-61.1
NN	SE	N	30	50.2 (48.9-51.6)	3.607614	45.3-58.8
NN	LO	N	30	51.7 (50.2-53.1)	3.828932	42.3-60.7
NN	HO	N	30	50.8 (49.4-52.1)	3.597886	44.0-58.6
LN	SE	N	30	53.8 (52.4-55.1)	3.721463	45.2-60.1
LN	LO	N	30	55.3 (53.6-57.1)	4.789029	43.8-65.6
LN	HO	N	30	55.3 (53.3-57.2)	5.296116	44.3-67.5
HN	SE	N	30	59.4 (57.5-61.4)	5.217251	44.9-68.8
HN	LO	N	30	60.8 (58.3-63.2)	6.478300	47.3-72.9
HN	HO	N	30	60.5 (58.0-63.0)	6.658969	47.7-73.0
NN	OE	Y	19	51.1 (49.0-53.2)	4.439369	43.1-61.9
NN	SE	Y	19	50.7 (48.4-53.1)	4.879689	43.5-59.2
NN	LO	Y	19	51.1 (49.0-53.2)	4.348249	43.1-59.8
NN	HO	Y	19	51.5 (49.3-53.8)	4.668555	44.2-60.4
LN	SE	Y	19	54.8 (52.6-56.9)	4.486175	48.4-61.6
LN	LO	Y	19	54.8 (53.1-56.5)	3.583020	47.3-59.8
LN	HO	Y	19	55.1 (52.7-57.4)	4.820974	48.1-66.4
HN	SE	Y	19	59.5 (57.2-61.8)	4.754466	51.2-68.4
HN	LO	Y	19	59.4 (57.1-61.6)	4.674263	52.3-67.3
HN	HO	Y	19	58.8 (56.2-61.4)	5.403664	51.4-72.2

Table-A II-2 Descriptive statistics with PTA grouping

Noise	Occlusion	Hearing Group	N	Mean (95% CI)	SD	Range
NN	OE	N	40	50.4 (49.2-51.7)	3.817930	43.1-61.1
NN	SE	N	40	49.8 (48.6-51.0)	3.791776	43.5-58.8
NN	LO	N	40	51.1 (49.8-52.4)	4.039563	42.3-60.7
NN	HO	N	40	50.5 (49.3-51.7)	3.726755	44.0-58.6
LN	SE	N	40	53.5 (52.3-54.7)	3.744235	45.2-60.1
LN	LO	N	40	54.8 (53.3-56.2)	4.555254	43.8-65.6
LN	HO	N	40	54.8 (53.2-56.4)	4.958381	44.3-67.5
HN	SE	N	40	58.9 (57.2-60.5)	5.108261	44.9-68.8
HN	LO	N	40	59.9 (57.9-61.8)	6.069983	47.3-72.9
HN	HO	N	40	59.5 (57.5-61.6)	6.331910	47.7-73.0
NN	OE	Y	9	53.2 (49.7-56.6)	4.528422	46.8-61.9
NN	SE	Y	9	53.3 (50.0-56.7)	4.390194	47.1-59.2
NN	LO	Y	9	53.2 (50.5-55.9)	3.512360	48.2-59.8
NN	HO	Y	9	53.5 (50.0-57.0)	4.589495	47.6-60.4
LN	SE	Y	9	57.0 (53.8-60.2)	4.173899	50.7-61.6
LN	LO	Y	9	56.7 (54.6-58.9)	2.800469	51.7-59.8
LN	HO	Y	9	56.8 (52.6-61.1)	5.514001	48.1-66.4
HN	SE	Y	9	62.0 (59.2-64.8)	3.647370	56.9-68.4
HN	LO	Y	9	61.7 (58.2-65.3)	4.610092	55.5-67.3
HN	HO	Y	9	61.3 (56.9-65.6)	5.680874	53.3-72.2

2. LME model output

Best model output for the effects of noise, ear occlusion, and hearing grouping:

```
Linear mixed model fit by REML.
t-tests use Satterthwaite's method [lmerModLmerTest]
Formula:
overall_SPL_dBA ~ noise + occlusion + HI_pta
+ (1 + noise + occlusion |  participant) + (1 | set)
Data: dt.spl %>%
    filter(occlusion != "HAOE", occlusion != "OE")

REML criterion at convergence: 1977.1
```

```
Scaled residuals:
Min   1Q  Median   3Q  Max
-3.1831 -0.4693  0.0117  0.4798  4.4586

Random effects:
 Groups   Name     Variance Std.Dev. Corr
 participant (Intercept) 16.60495 4.0749
             noise.L   7.41523 2.7231 0.52
             noise.Q   0.08792 0.2965 0.26  0.80
             occlusion.L  2.39660 1.5481 0.43  0.55  0.88
          occlusion.Q  1.23254 1.1102   -0.17 -0.52 -0.24  0.02
 set      (Intercept)  0.07972 0.2823
 Residual              1.95588 1.3985
Number of obs: 441, groups:  participant, 49; set, 10

Fixed effects:
         Estimate Std. Error    df t value Pr(>|t|)
(Intercept)  54.6360  0.6381  52.1898  85.626  < 2e-16 ***
noise.L       6.2457  0.4058  47.9391  15.391  < 2e-16 ***
noise.Q       0.4658  0.1230 111.1292   3.787 0.000248 ***
occlusion.L   0.5008  0.2510  48.6119   1.995 0.051642 .
occlusion.Q  -0.4874  0.1975  48.4508  -2.468 0.017149 *
HI_ptaY       3.1419  1.2867  46.9045   2.442 0.018441 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
         (Intr) nois.L nois.Q occl.L occl.Q
```

```
noise.L      0.454
noise.Q      0.082  0.264
occlusion.L  0.342  0.460  0.266
occlusion.Q -0.128 -0.397 -0.066  0.011
HI_ptaY     -0.370  0.000  0.000  0.001 -0.001
optimizer (nloptwrap) convergence code: 0 (OK)
```

BIBLIOGRAPHY

Ahrens, A. & Lund, K. D. (2022). Auditory spatial analysis in reverberant multi-talker environments with congruent and incongruent audio-visual room information. *The Journal of the Acoustical Society of America*, 152(3), 1586–1594. doi: 10.1121/10.0013991.

American National Standards Institute. (2008). *Methods for Measuring the Real-Ear Attenuation of Hearing Protectors* (Report n°ANSI/ASA S12.6-2008).

Azadi, H., Akbarzadeh-T, M.-R., Shoeibi, A., Kobravi, H. R. et al. (2021). Evaluating the effect of Parkinson's disease on jitter and shimmer speech features. *Advanced Biomedical Research*, 10(1), 54.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi: 10.18637/jss.v067.i01.

Bellows, S. & Leishman, T. W. (2022, September). Effect of Head Orientation on Speech Directivity. *Interspeech 2022*, pp. 246–250. doi: 10.21437/Interspeech.2022-553.

Black, J. W. (1950). The Effect of Room Characteristics upon Vocal Intensity and Rate. *The Journal of the Acoustical Society of America*, 22(2), 174–176. doi: 10.1121/1.1906585.

Boersma, P. & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot international*, 5, 341–345.

Botelho, C., Abad, A., Schultz, T. & Trancoso, I. (2024). Speech as a Biomarker for Disease Detection. *IEEE Access*, 12, 184487–184508. doi: 10.1109/ACCESS.2024.3506433.

Bottalico, P. (2018). Lombard effect, ambient noise, and willingness to spend time and money in a restaurant. *The Journal of the Acoustical Society of America*, 144(3), EL209–EL214. doi: 10.1121/1.5055018.

Bottalico, P. & Murgia, S. (2023). The Effect of the Frequency and Energetic Content of Broadband Noise on the Lombard Effect and Speech Intelligibility. *Acoustics*, 5(4), 898–908. doi: 10.3390/acoustics5040052.

Bottalico, P., Graetzer, S. & Hunter, E. J. (2016). Effects of speech style, room acoustics, and vocal fatigue on vocal effort. *The Journal of the Acoustical Society of America*, 139(5), 2870–2879. doi: 10.1121/1.4950812.

Bottalico, P., Passione, I. I., Graetzer, S. & Hunter, E. (2017). Evaluation of the Starting Point of the Lombard Effect. *Acta Acustica united with Acustica*, 103(1), 169–172. doi: 10.3813/aaa.919043. Publisher: European Acoustics Association.

Bouserhal, R. E., Bockstael, A., MacDonald, E., Falk, T. H. & Voix, J. (2017a). Modeling Speech Level as a Function of Background Noise Level and Talker-to-Listener Distance for Talkers Wearing Hearing Protection Devices. *Journal of Speech Language and Hearing Research*, 60(12), 3393. doi: 10.1044/2017_JSLHR-S-17-0052.

Bouserhal, R. E., Falk, T. H. & Voix, J. (2017b). In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension. *The Journal of the Acoustical Society of America*, 141(3), 1321–1331. doi: 10.1121/1.4976051.

Bouserhal, R. E., Bernier, A. & Voix, J. (2019). An in-ear speech database in varying conditions of the audio-phonation loop. *The Journal of the Acoustical Society of America*, 145(2), 1069–1077. doi: 10.1121/1.5091777.

Brajot, F.-X., Nguyen, D., DiGiovanni, J. & Gracco, V. L. (2018). The impact of perilaryngeal vibration on the self-perception of loudness and the Lombard effect. *Experimental Brain Research*, 236(6), 1713–1723. doi: 10.1007/s00221-018-5248-9.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press. doi: 10.7551/mitpress/1486.001.0001.

Brockmann-Bauser, M., Bohlender, J. & Mehta, D. (2018). Acoustic Perturbation Measures Improve with Increasing Vocal Intensity in Individuals With and Without Voice Disorders. *Journal of Voice*, 32(2), 162–168. doi: 10.1016/j.jvoice.2017.04.008. Publisher: Elsevier BV.

Brown, J. & Sonderegger, M. (2024, June). Creaky voice variation across language, gender and age in Canadian English-French bilingual speech. *LabPhon19*.

Brumm, H. & Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour*, 148(11/13), 1173–1198. Retrieved from: https://www.jstor.org/stable/41445240. Publisher: BRILL.

Byeon, H. (2021). Associations between adolescents' earphone usage in noisy environments, hearing loss, and self-reported hearing problems in a nationally representative sample of South Korean middle and high school students. *Medicine*, 100(3), e24056. doi: 10.1097/MD.0000000000024056.

Cai, X., Yin, Y. & Zhang, Q. (2020). A cross-language study on feedforward and feedback control of voice intensity in Chinese–English bilinguals. *Applied Psycholinguistics*, 41(4), 771–795. doi: 10.1017/S0142716420000223.

Carillo, K., Doutres, O. & Sgard, F. (2021). Numerical investigation of the earplug contribution to the low-frequency objective occlusion effect induced by bone-conducted stimulation. *The Journal of the Acoustical Society of America*, 150(3), 2006–2023. doi: 10.1121/10.0006209. Publisher: Acoustical Society of America (ASA).

Carle, R., Laugesen, S. & Nielsen, C. (2002). Observations on the Relations among Occlusion Effect, Compliance, and Vent Size. *Journal of the American Academy of Audiology*, 13(01), 025–037. doi: 10.1055/s-0040-1715945. Publisher: American Academy of Audiology.

Chabot, P., Bouserhal, R. E., Cardinal, P. & Voix, J. (2021). Detection and classification of human-produced nonverbal audio events. *Applied Acoustics*, 171, 107643. doi: 10.1016/j.apacoust.2020.107643.

Chadha, S., Kamenov, K. & Cieza, A. (2021). The world report on hearing, 2021. *Bulletin of the World Health Organization*, 99, 242 - 242A. Retrieved from: https://api.semanticscholar.org/CorpusID:233581663.

Cipriano, M., Astolfi, A. & Pelegrín-García, D. (2017). Combined effect of noise and room acoustics on vocal effort in simulated classrooms. *The Journal of the Acoustical Society of America*, 141(1), EL51–EL56. doi: 10.1121/1.4973849.

Clark, J. G. (1981). Uses and Abuses of Hearing Loss Classification. *Asha*, 23, 493–500.

Coelho, A. C., Medved, D. M. & Brasolotto, A. G. (2015). Hearing Loss and the Voice. In Bahmad, F. (Ed.), *Update On Hearing Loss*. InTech. doi: 10.5772/61217.

Cowie, R. & Douglas-Cowie, E. (1992). *Postlingually Acquired Deafness: Speech Deterioration and the Wider Consequences*. DE GRUYTER MOUTON. doi: 10.1515/9783110869125.

Daşdöğen, Ü., Awan, S. N., Bottalico, P., Iglesias, A., Getchell, N. & Abbott, K. V. (2023). The Influence of Multisensory Input On Voice Perception and Production Using Immersive Virtual Reality. *Journal of Voice*, S0892199723002357. doi: 10.1016/j.jvoice.2023.07.026.

Dalebout, S. (2009). *The Praeger guide to hearing and hearing loss: assessment, treatment, and prevention*. Westport, Conn: Praeger Publishers.

Danhauer, J. L., Johnson, C. E., Byrd, A., DeGood, L., Meuel, C., Pecile, A. & Koch, L. L. (2009). Survey of College Students on iPod Use and Hearing Health. *Journal of the American Academy of Audiology*, 20(1), 5–27. doi: 10.3766/jaaa.20.1.2.

Defays, A., Safin, S., Billon, A., Decaestecker, C., Warzée, N., Leclercq, P. & Nyssen, A.-S. (2015). Bimodal Interaction: The Role of Visual Information in Performing Acoustic Assessment in Architecture. *The Ergonomics Open Journal*, 7(1), 13–20. doi: 10.2174/1875934301407010013.

Dehankar, S. S. & Gaurkar, S. S. (2022). Impact on Hearing Due to Prolonged Use of Audio Devices: A Literature Review. *Cureus*. doi: 10.7759/cureus.31425.

Deliyski, D. D., Shaw, H. S., Evans, M. K. & Vesselinov, R. (2006). Regression tree approach to studying factors influencing acoustic voice analysis. *Folia Phoniatrica et Logopaedica*, 58(4), 274–288. doi: 10.1159/000093184.

Denk, F., Hieke, T., Roberz, M. & Husstedt, H. (2023). Occlusion and coupling effects with different earmold designs – all a matter of opening the ear canal? *International Journal of Audiology*, 62(3), 227–237. doi: 10.1080/14992027.2022.2039966. Publisher: Informa UK Limited.

Di Stadio, A., Sossamon, J., De Luca, P., Indovina, I., Motta, G., Ralli, M., Brenner, M. J., Frohman, E. M. & Plant, G. T. (2025). "Do You Hear What I Hear?" Speech and Voice Alterations in Hearing Loss: A Systematic Review. *Journal of Clinical Medicine*, 14(5), 1428. doi: 10.3390/jcm14051428. Publisher: MDPI AG.

Dunn, H. K. & Farnsworth, D. W. (1939). Exploration of Pressure Field Around the Human Head During Speech. *The Journal of the Acoustical Society of America*, 10(3), 184–199. doi: 10.1121/1.1915975.

Ebbutt, N., Shamei, A., Purnomo, C. & Gick, B. (2021). Prosodic characteristics of English speakers with Alzheimer's disease. *The Journal of the Acoustical Society of America*, 150(4_Supplement), A274–A274. doi: 10.1121/10.0008273.

Feder, K., Michaud, D., McNamee, J., Fitzpatrick, E., Davies, H. & Leroux, T. (2017). Prevalence of Hazardous Occupational Noise Exposure, Hearing Loss, and Hearing Protection Usage Among a Representative Sample of Working Canadians. *Journal of Occupational & Environmental Medicine*, 59(1), 92–113. doi: 10.1097/JOM.0000000000000920.

Feder, K., Marro, L., McNamee, J. & Michaud, D. (2019). Prevalence of loud leisure noise activities among a representative sample of Canadians aged 6–79 years. *The Journal of the Acoustical Society of America*, 146(5), 3934–3946. doi: 10.1121/1.5132949.

Fox, J. & Weisberg, S. (2019). *An R Companion to Applied Regression* (ed. 3rd). Thousand Oaks CA: Sage. Retrieved from: https://www.john-fox.ca/Companion/index.html.

Fraile, R. & Godino-Llorente, J. I. (2014). Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control*, 14, 42–54. doi: 10.1016/j.bspc.2014.07.001.

Garnier, M. & Henrich, N. (2014). Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech & Language*, 28(2), 580–597. doi: https://doi.org/10.1016/j.csl.2013.07.005.

Garnier, M., Henrich, N. & Dubois, D. (2010). Influence of Sound Immersion and Communicative Interaction on the Lombard Effect. *Journal of Speech, Language and Hearing Research*, 53(3), 588–608. doi: 10.1044/1092-4388(2009/08-0138).

Garnier, M., Ménard, L. & Alexandre, B. (2018). Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues? *The Journal of the Acoustical Society of America*, 144(2), 1059–1074. doi: 10.1121/1.5051321.

Gick, B., Schellenberg, M., Stavness, I. & Taylor, R. C. (2019). *Articulatory phonetics* (ed. 1). Routledge.

Giguère, C., Laroche, C., Brault, E., Ste-Marie, J.-C., Brosseau-Villeneuve, M., Philippon, B. & Vaillancourt, V. (2006). Quantifying the Lombard effect in different background noises. *The Journal of the Acoustical Society of America*, 120(5_Supplement), 3378–3378. doi: 10.1121/1.4781616. Publisher: Acoustical Society of America (ASA).

Godoy, E., Koutsogiannaki, M. & Stylianou, Y. (2014). Approaching speech intelligibility enhancement with inspiration from Lombard and Clear speaking styles. *Computer Speech & Language*, 28(2), 629–647. doi: 10.1016/j.csl.2013.09.007.

Goman, A. M. & Lin, F. R. (2016). Prevalence of Hearing Loss by Severity in the United States. *American journal of public health*, 106 10, 1820-2. Retrieved from: https://api.semanticscholar.org/CorpusID:207278241.

Goverdovsky, V., Von Rosenberg, W., Nakamura, T., Looney, D., Sharp, D. J., Papavassiliou, C., Morrell, M. J. & Mandic, D. P. (2017). Hearables: Multimodal physiological in-ear sensing. *Scientific Reports*, 7(1), 6948. doi: 10.1038/s41598-017-06925-2.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39(5), 350–365. doi: https://doi.org/10.1016/j.jcomdis.2006.06.013.

Hansen, J. H. L., Lee, J., Ali, H. & Saba, J. N. (2020). A speech perturbation strategy based on "Lombard effect" for enhanced intelligibility for cochlear implant listeners. *The Journal of the Acoustical Society of America*, 147(3), 1418–1428. doi: 10.1121/10.0000690.

Hansen, M. (1997). *Occlusion effects, Part I*. (Ph.D. thesis, Technical University of Denmark).

Hauret, J., Olivier, M., Joubaud, T., Langrenne, C., Poirée, S., Zimpfer, V. & Bavu, E. (2025). Vibravox: A dataset of french speech captured with body-conduction audio sensors. *Speech Communication*, 172, 103238. doi: 10.1016/j.specom.2025.103238. Publisher: Elsevier BV.

Hollien, H. (1974). On vocal registers. *Journal of Phonetics*, 2(2), 125–143. doi: 10.1016/S0095-4470(19)31188-X.

Hormann, H., Lazarus-Mainka, G., Schubeius, M. & Lazarus, H. (1984). The Effects of Noise and the Wearing of Ear Protectors on Verbal Communication. *Noise Control Engineering Journal*, 23(2), 69–77.

Houde, J. F. & Nagarajan, S. S. (2011). Speech Production as State Feedback Control. *Frontiers in Human Neuroscience*, 5. doi: 10.3389/fnhum.2011.00082.

International Organization for Standardization. (1994). *Estimation of effective A-weighted sound pressure levels when hearing protectors are worn*.

Jiménez-Jiménez, F. J., Gamboa, J., Nieto, A., Guerrero, J., Orti-Pareja, M., Molina, J. A., García-Albea, E. & Cobeta, I. (1997). Acoustic voice analysis in untreated patients with Parkinson's disease. *Parkinsonism & Related Disorders*, 3(2), 111–116.

Junqua, J.-C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication*, 20(1-2), 13–22. doi: 10.1016/S0167-6393(96)00041-6.

Kato, H., Takemoto, H., Nishimura, R. & Mokhtari, P. (2010). Spatial acoustic cues for the auditory perception of speaker's facing direction. *Proceedings of the 20th International Congress on Acoustics (ICA 2010)*. Retrieved from: https://api.semanticscholar.org/CorpusID:7464355.

Kryter, K. D. (1946). Effects of Ear Protective Devices on the Intelligibility of Speech in Noise. *Journal of the Acoustical Society of America*, 18, 413–417.

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82, 1–26. doi: 10.18637/jss.v082.i13.

Lane, H. & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech, Language and Hearing Research*, 14(4), 677. Retrieved from: http://jslhr.asha.org/cgi/content/abstract/14/4/677. 00569.

Lane, H. & Webster, J. W. (1991). Speech deterioration in postlingually deafened adults. *The Journal of the Acoustical Society of America*, 89(2), 859–866. doi: 10.1121/1.1894647.

Lee, G.-S. (2012). Variability in Voice Fundamental Frequency of Sustained Vowels in Speakers With Sensorineural Hearing Loss. *Journal of Voice*, 26(1), 24–29. doi: 10.1016/j.jvoice.2010.10.003. Publisher: Elsevier BV.

Leishman, T. W., Bellows, S. D., Pincock, C. M. & Whiting, J. K. (2021). High-resolution spherical directivity of live speech from a multiple-capture transfer function method. *The Journal of the Acoustical Society of America*, 149(3), 1507–1523. doi: 10.1121/10.0003363.

Lenth, R. V. [R package version 1.7.5]. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means. Retrieved from: https://CRAN.R-project.org/package=emmeans.

Levelt, W. J. M., Roelofs, A. & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1), 1–38. doi: 10.1017/s0140525x99001776. Publisher: Cambridge University Press (CUP).

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312. Retrieved from: http://arxiv.org/abs/1405.0312.

Lombard, E. (1911). Le signe de l'élévation de la voix. *Annales des Maladies de l'Oreille, du Larynx, du Nez et du Pharynx*, 37, 101–119. English translation: "The sign of the elevation of the voice".

Lu, Y. & Cooke, M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, 51(12), 1253–1262. doi: 10.1016/j.specom.2009.07.002.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017, August). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech 2017*, pp. 498–502. doi: 10.21437/Interspeech.2017-1386.

Mehrban, M. H., Voix, J. & Bouserhal, R. E. (2024). Classification of breathing phase and path with in-ear microphones. *Sensors*, 24(20), 6679.

Meinke, D. K., Berger, E. H., Driscoll, D. P., Neitzel, R. L. & Bright, K. (Eds.). (2022). *The noise manual* (ed. 6th edition). Falls Church, VA: American Industrial Hygiene Association.

Mitsuya, T. & Purcell, D. W. (2016). Occlusion effect on compensatory formant production and voice amplitude in response to real-time perturbation. *The Journal of the Acoustical Society of America*, 140(6), 4017–4026. doi: 10.1121/1.4968539.

Moriarty, B. T., Ananthanarayana, R. M. & Monson, B. B. (2024). Factors influencing the minimum audible change in talker head orientation cues using diotic stimuli. *The Journal of the Acoustical Society of America*, 156(2), 763–773. doi: 10.1121/10.0028119.

Nagy, A., Elshafei, R. & Mahmoud, S. (2020). Correlating Undiagnosed Hearing Impairment with Hyperfunctional Dysphonia. *Journal of Voice*, 34(4), 616–621. doi: 10.1016/j.jvoice.2019.02.002.

Navarro, R. (1996). Effects of Ear Canal Occlusion and Masking on the Perception of Voice. *Perceptual and Motor Skills*, 82(1), 199–208. doi: 10.2466/pms.1996.82.1.199. PMID: 8668476.

Nijs, L., Saher, K. & Den Ouden, D. (2008). Effect of room absorption on human vocal output in multitalker situations. *The Journal of the Acoustical Society of America*, 123(2), 803–813. doi: 10.1121/1.2821410. Publisher: Acoustical Society of America (ASA).

Nilsson, M., Soli, S. D. & Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2), 1085–1099. doi: 10.1121/1.408469.

Novitasari, S., Sakti, S. & Nakamura, S. (2022). A Machine Speech Chain Approach for Dynamically Adaptive Lombard TTS in Static and Dynamic Noise Environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 2673–2688. doi: 10.1109/TASLP.2022.3196879.

Nudelman, C. J. & Bottalico, P. (2025). Investigating the Impact of Visual Input on Voice Production in Virtual Reality. *Journal of Voice*, 39(4), 1053–1064. doi: https://doi.org/10.1016/j.jvoice.2023.07.016.

Parrell, B., Lammert, A. C., Ciccarelli, G. & Quatieri, T. F. (2019). Current models of speech motor control: A control-theoretic overview of architectures and properties. *The Journal of the Acoustical Society of America*, 145(3), 1456–1481. doi: 10.1121/1.5092807.

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Švec, J. G. & Hillman, R. (2018). Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *American Journal of Speech-Language Pathology*, 27(3), 887–905. doi: 10.1044/2018_AJSLP-17-0009.

Patel, R., Niziolek, C., Reilly, K. & Guenther, F. H. (2011). Prosodic Adaptations to Pitch Perturbation in Running Speech. *Journal of Speech, Language, and Hearing Research*, 54(4), 1051–1059. doi: 10.1044/1092-4388(2010/10-0162).

Pearsons, K. S., Bennett, R. L. & Fidell, S. (1977). *Speech Levels in Various Noise Environments* (Report n°EPA-600/1-77-025). Washington, DC.

Pelegrín-García, D., Smits, B., Brunskog, J. & Jeong, C. (2011a). Vocal effort with changing talker-to-listener distance in different acoustic environments. *The Journal of the Acoustical Society of America*, 129(4), 1981–1990. doi: 10.1121/1.3552881. 00039.

Pelegrín-García, D., Fuentes-Mendizábal, O., Brunskog, J. & Jeong, C.-H. (2011b). Equal autophonic level curves under different room acoustics conditions. *The Journal of the Acoustical Society of America*, 130(1), 228–238. doi: 10.1121/1.3598429. 00026.

Perkell, J., Lane, H., Svirsky, M. & Webster, J. (1992). Speech of cochlear implant patients: A longitudinal study of vowel production. *The Journal of the Acoustical Society of America*, 91(5), 2961–2978. doi: 10.1121/1.402932.

Pittman, A. L., Daliri, A. & Meadows, L. (2018). Vocal Biomarkers of Mild-to-Moderate Hearing Loss in Children and Adults: Voiceless Sibilants. *Journal of Speech, Language, and Hearing Research*, 61(11), 2814–2826. doi: 10.1044/2018_JSLHR-H-17-0460.

Postma, A. (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition*, 77(2), 97–132. doi: 10.1016/s0010-0277(00)00090-1. Publisher: Elsevier BV.

Purcell, D. W. & Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America*, 119(4), 2288–2297. doi: 10.1121/1.2173514.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.

Ramage-Morin, P. L., Banks, R., Pineault, D. & Atrach, M. (2019, August 21). Unperceived Hearing Loss among Canadians Aged 40 to 79. Statistics Canada. Retrieved from: https://www150.statcan.gc.ca/n1/pub/82-003-x/2019008/article/00002-eng.htm.

Röddiger, T., Clarke, C., Breitling, P., Schneegans, T., Zhao, H., Gellersen, H. & Beigl, M. (2022). Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(3), 1–57. doi: 10.1145/3550314.

Saint-Gaudens, H., Nélisse, H., Sgard, F. & Doutres, O. (2022). Towards a practical methodology for assessment of the objective occlusion effect induced by earplugs. *The Journal of the Acoustical Society of America*, 151(6), 4086–4100. doi: 10.1121/10.0011696.

Saltzman, E. L. & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology*, 1(4), 333–382. doi: 10.1207/s15326969eco0104_2.

Schutte, M., Ewert, S. D. & Wiegrebe, L. (2019). The percept of reverberation is not affected by visual room impression in virtual environments. *The Journal of the Acoustical Society of America*, 145(3), EL229–EL235. doi: 10.1121/1.5093642.

Seol, H. Y. & Moon, I. J. (2022). Hearables as a Gateway to Hearing Health Care: A Review. *Clinical and Experimental Otorhinolaryngology*, 15(2), 127–134. doi: 10.21053/ceo.2021.01662.

Shamei, A., Liu, Y. & Gick, B. (2023). Reduction of vowel space in Alzheimer's disease. *JASA Express Letters*, 3(3), 035202. doi: 10.1121/10.0017438.

Siegel, G. M. & Pick, H. L. (1974). Auditory feedback in the regulation of voice. *The Journal of the Acoustical Society of America*, 56(5), 1618–1624. doi: 10.1121/1.1903486.

Sierra-Polanco, T., Cantor-Cutiva, L. C., Hunter, E. J. & Bottalico, P. (2021). Changes of Voice Production in Artificial Acoustic Environments. *Frontiers in Built Environment*, 7, 666152. doi: 10.3389/fbuil.2021.666152.

Simpson, G. L. [R package version 0.9.2]. (2024). gratia: Graceful ggplot-Based Graphics and Other Functions for GAMs Fitted using mgcv. Retrieved from: https://gavinsimpson.github.io/gratia/.

Skodda, S., Visser, W. & Schlegel, U. (2011). Vowel Articulation in Parkinson's Disease. *Journal of Voice*, 25(4), 467–472. doi: 10.1016/j.jvoice.2010.01.009.

Solana-Lavalle, G., Galán-Hernández, J.-C. & Rosas-Romero, R. (2020). Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. *Biocybernetics and Biomedical Engineering*, 40(1), 505–516.

Statistics Canada. [Archived version; published by the authority of the Minister responsible for Statistics Canada, Ottawa]. (2021, October 20). Hearing health of Canadian adults. Retrieved from: https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2021077-eng.htm.

Styler, W. (2023). Using Praat for Linguistic Research. Retrieved from: http://savethevowels.org/praat.

Sørensen, A. J. M., Lunner, T. & MacDonald, E. N. (2024). Conversational Dynamics in Task Dialogue Between Interlocutors With and Without Hearing Impairment. *Trends in Hearing*, 28. doi: 10.1177/23312165241296073. Publisher: SAGE Publications.

Tatham, M. & Morton, K. (2007). *Speech production and perception* (ed. Nachdr.). Basingstoke, Hampshire: Palgrave Macmillan.

Team, R. C. (2022). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: https://www.R-project.org/.

The MathWorks Inc. (2022). MATLAB (Version 9.13.0 (R2022b)). Natick, Massachusetts, United States: The MathWorks Inc. Retrieved from: https://www.mathworks.com.

Tomassi, N. E., Castro, M. E., Timmons Sund, L., Díaz-Cádiz, M. E., Buckley†, D. P. & Stepp, C. E. (2023). Effects of Sidetone Amplification on Vocal Function During Telecommunication. *Journal of Voice*, 37(4), 553–560. doi: 10.1016/j.jvoice.2021.03.027.

Tremblay, S., Shiller, D. M. & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature*, 423(6942), 866–869. doi: 10.1038/nature01710.

Tsang, K. Y. & Mannion, D. J. (2022). Relating Sound and Sight in Simulated Environments. *Multisensory Research*, 35(7-8), 589–622. doi: 10.1163/22134808-bja10082.

Tsang, K. Y.-T. (2023). *The influence of vision on the perceptual compensation for reverberation in simulated environments*. (Ph.D. thesis, [object Object]). Retrieved from: http://hdl.handle.net/1959.4/101368.

Tufts, J. B. & Frank, T. (2003). Speech production in noise with and without hearing protection. *The Journal of the Acoustical Society of America*, 114(2), 1069. doi: 10.1121/1.1592165.00032.

Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A., Soli, S. D. & Giguère, C. (2005). Adaptation of the HINT (hearing in noise test) for adult Canadian Francophone populations. *International Journal of Audiology*, 44(6), 358–361. doi: 10.1080/14992020500060875.

van Rij, J., Wieling, M., Baayen, R. H. & van Rijn, H. [R package version 2.4.1]. (2022). itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.

Vaziri, G., Giguère, C. & Dajani, H. R. (2022). The effect of hearing protection worn by talker and/or target listener on speech production in quiet and noise. *The Journal of the Acoustical Society of America*, 152(3), 1528–1538. doi: 10.1121/10.0013895.

Villacorta, V. M., Perkell, J. S. & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America*, 122(4), 2306–2319. doi: 10.1121/1.2773966.

Villegas, J., Perkins, J. & Wilson, I. (2021). Effects of task and language nativeness on the Lombard effect and on its onset and offset timing. *The Journal of the Acoustical Society of America*, 149(3), 1855–1865. doi: 10.1121/10.0003772.

Vogel, A. P. & Morgan, A. T. (2009). Factors affecting the quality of sound recording for speech and voice analysis. *International Journal of Speech-Language Pathology*, 11(6), 431–437. doi: 10.3109/17549500902822189.

Voix, J. & Laville, F. (2009). The objective measurement of individual earplug field performance. *The Journal of the Acoustical Society of America*, 125(6), 3722–3732. doi: 10.1121/1.3125769.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from: https://ggplot2.tidyverse.org.

Wickham, H., François, R., Henry, L. & Müller, K. [R package version 1.0.9]. (2022). dplyr: A Grammar of Data Manipulation. Retrieved from: https://CRAN.R-project.org/package=dplyr.

Winkler, A., Latzel, M. & Holube, I. (2016). Open Versus Closed Hearing-Aid Fittings: A Literature Review of Both Fitting Approaches. *Trends in Hearing*, 20. doi: 10.1177/2331216516631741. Publisher: SAGE Publications.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R* (ed. 2). Chapman and Hall/CRC.

Xu, Y., Larson, C. R., Bauer, J. J. & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *The Journal of the Acoustical Society of America*, 116(2), 1168–1178. doi: 10.1121/1.1763952.

Yu, J.-F., Lee, K.-C., Wang, R.-H., Chen, Y.-S., Fan, C.-C., Peng, Y.-C., Tu, T.-H., Chen, C.-I. & Lin, K.-Y. (2015). Anthropometry of external auditory canal by non-contactable measurement. *Applied Ergonomics*, 50, 50–55. doi: 10.1016/j.apergo.2015.01.008.

Zelcer, S., Henri, C., Tewfik, T. L. & Mazer, B. (2002). Multidimensional voice program analysis (MDVP) and the diagnosis of pediatric vocal cord dysfunction. *Annals of Allergy, Asthma & Immunology*, 88(6), 601–608. doi: 10.1016/S1081-1206(10)61892-3.

Zhang, X., Verduyckt, I. & Bouserhal, R. (2021). Speech production for hearing impaired talkers in noise with ear occlusion. *Canadian Acoustics*, 49(3), 71. Retrieved from: https://awc.caa-aca.ca/index.php/AWC/AWC21/paper/view/787.

Zhang, X., Clayards, M. & Bouserhal, R. (2022, August). The effect of vowel lengthening on the intelligibility of occluded Lombard speech. *Canadian Acoustics*, 50(3), 86–87. Retrieved from: https://jcaa.caa-aca.ca/index.php/jcaa/article/view/3884.

Zhang, X., Berger, L. E., Tran, D.-H. & Bouserhal, R. (2023, October). Enhancing Automatic Speech Recognition of a Regional Dialect: A Pilot Study with Québécois French. *Canadian Acoustics*, 51(3), 76–77. Retrieved from: https://jcaa.caa-aca.ca/index.php/jcaa/article/view/4117.

Zhang, X., Braga, A., Shamei, A. & Bouserhal, R. (2024a). A comparison of voice quality measures across in-ear, outer-ear, and standard microphone recordings. *The Journal of the Acoustical Society of America*, 156(4_Supplement), A55–A55. doi: 10.1121/10.0035092.

Zhang, X., Shamei, A., Grond, F., Verduyckt, I. & Bouserhal, R. (2024b). Towards a better understanding of multimodal integration and sensorimotor adaptation to audiovisual environmental incongruence using Virtual Reality. *The Journal of the Acoustical Society of America*, 155(3_Supplement), A178–A179. doi: 10.1121/10.0027233.

Zhang, X., Verduyckt, I. & Bouserhal, R. (2025, May). Speech Production in Challenging Listening Conditions. *Proceedings of the 7th CIRMMT-OICRM-BRAMS Student Colloquium*. Retrieved from: https://www.ccob-cobs.org/2025/oral-presentations#xinyi-zhang.

Švec, J. G. & Granqvist, S. (2010). Guidelines for Selecting Microphones for Human Voice Production Research. *American Journal of Speech-Language Pathology*, 19(4), 356–368. doi: 10.1044/1058-0360(2010/09-0091).