I Know You're Listening: Adaptive Voice for HRI

by

Paige Irène Tuttösí

B.Sc., Simon Fraser University, 2021
B.A. (Hons.), Simon Fraser University, 2016

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Spring 2025

Declaration of Committee

Name: **Paige Irène Tuttösí**

Degree: **Doctor of Philosophy**

Thesis title: **I Know You're Listening: Adaptive Voice for HRI**

Committee:

Chair: Anders Miltner
Assistant Professor, Computing Science

Angelica Lim
Supervisor
Assistant Professor, Computing Science

Jean-Julien Aucouturier
Committee Member
Directeur de Recherche, FEMTO-ST Institute
Centre national de la recherche scientifique (CNRS)
Besançon France

Mo Chen
Committee Member
Associate Professor, Computing Science

Lawrence Kim
Examiner
Assistant Professor, Computing Science

Tony Belpaeme
External Examiner
Professor, Faculty of Engineering and Architecture
Ghent University
Ghent Belgium

Abstract

With increased globalization, the desire to learn a new language is putting increased pressure on the need for second language (L2) teachers. Social robots are particularly well suited for teaching tasks as they can create engaging social interactions in a physical form, for example, using action stories (i.e., acting out an action matching the word) when learning new vocabulary. While the use of social robots for language teaching has been explored, there remains limited work on a task-specific synthesized voices for language teaching robots. Given that language is a verbal task, this gap may have severe consequences for the effectiveness of robots for language teaching tasks. We address this lack of L2 teaching robot voices through three contributions: 1. We address the need for a lightweight and expressive robot voice. Using a fine-tuned version of Matcha-TTS, we use emoji prompting to create an expressive voice that shows a range of expressivity over time. The voice can run in real time with limited compute resources. Through case studies, we found this voice more expressive, socially appropriate, and suitable for long periods of expressive speech, such as storytelling. 2. We explore how to adapt a robot's voice to physical and social ambient environments to deploy our voices in various locations. We found that increasing pitch and pitch rate in noisy and high-energy environments makes the robot's voice appear more appropriate and makes it seem more aware of its current environment. 3. We create an English TTS system with improved clarity for L2 listeners using known linguistic properties of vowels that are difficult for these listeners. We used a data-driven, perception-based approach to understand how L2 speakers use duration cues to interpret challenging words with minimal tense (long) and lax (short) vowels in English. We found that the duration of vowels strongly influences the perception for L2 listeners and created an "L2 clarity mode" for Matcha-TTS that applies a lengthening to tense vowels while leaving lax vowels unchanged. Our clarity mode was found to be more respectful, intelligible, and encouraging than base Matcha-TTS while reducing transcription errors in these challenging tense/lax minimal pairs.

Keywords: social robotics; L2 teaching robots; L2-tailored TTS; adaptive speech synthesis; L2 speech perception; accessible speech synthesis; expressive speech synthesis; ambient appropriate voice

Dedication

This thesis is dedicated to my soul, my feelings of hope and joy, and any love that I once had for academia and research. You were once great friends who allowed me to have a sense of purpose in my life, but sadly you did not survive the long and harrowing journey of my doctoral studies. You are greatly missed and I write this thesis in your loving memory.

Acknowledgements

First, I would like to thank my dogs Ilona and Nymuë, who put up with missing walks for me to work all the time but also keep me sane. I want to thank my cat, Lily. She didn't make the move to France, but she has literally been with me for the entire 15 years of my university career and has put up with all of it. It may sound silly, but really, they all kept me going through all of this.

I want to thank my supervisor, Dr. Angelica Lim. I know so many people who have had awful experiences with their graduate studies, getting in fights and not feeling supported... I may have had some awful experiences, but not a single one of them could ever be due to Angelica. She is incredibly supportive, which is probably hard to do with me because I can be stubborn, thinking I know the best way to do something. I wanted to do crazy things, change my mind about what I was doing, do tons of work in linguistics, leave for France, and not come back. She supported me and not only supported but championed me through all of it. Without her, I would not have my connections, my job, or anything I have coming out of my PhD. She is really everything a supervisor should be, and I wish everyone could have this kind of experience for their graduate studies. She works so hard to be an amazing researcher, teacher, woman in STEM activist, supervisor, mother to her daughter, and second mother to all her students (she probably worried more about me getting a gun waved at me at a conference in Detroit than my own mother did). I am so lucky to not only have her as my supervisor but also to have her in my life in general.

I would also like to thank Dr. JJ Aucouturier, who allowed me to invade his lab and his country and never leave. Letting me argue with him about how I think things need to be a certain way and letting me complain about French bureaucracy. You have been such a wealth of knowledge in a domain I knew hardly anything about before I arrived here, and that help has been immeasurably beneficial to my PhD. I want to thank Dr. Mo Chen, who, despite being from a discipline slightly removed from my own, has been here throughout my PhD, giving me helpful feedback and direction along the way. Thank you to Dr. Henny Yeung and Dr. Yue Wang for giving me such excellent feedback and providing references and guidance as I stumbled my way through learning about linguistics.

I want to thank Shivam for allowing me to message him incessantly about Matcha-TTS and TTS in general. He is the reason why I can even call myself a TTS researcher now; I have learned so much from his help. I would like to thank Emma for letting me help her with

her ambiance paper; this basically kickstarted me knowing what I wanted to do with my PhD. Also, thank you for the general conversation and venting; it has been pretty... pretty... pretty good. I want to thank all my other Rosie lab mates, Bita, Shay, Bermet, Yasaman, Zhitian, Payam, Boey and Zach, and Aynaz, and Zhenxing in Neuro-Group who have helped me with projects, figuring out how to sort out my thesis, use solar and a myriad of other support. A special shout out to Rudra, who has helped to leave the lab early and give my dogs medication so I can go into Paris for work. Thank you to my colleagues, Dr. Thomas Janssoone, Dr. Mohammed Hafsati, Dr. Waldez Gomes, and Lucas, who have helped me learn how to integrate my work into our robots, given me helpful feedback, and helped me run my studies. I also want to thank Dr. Anita Tino, who is the only reason I made it as far as I did in engineering and helped me see how to be a great teacher for challenging topics. Also, a thank you to the people in my past lives, Dr. Hugo Cardoso and Shannon Wood. Even though I left archaeology behind, the 8 years I spent digging in jars of teeth cataloging human remains for repatriation has profoundly impacted the rest of my career. There is also a plethora of people in the HRI and speech communities that have given me useful support, feedback and direction, and I am sure no matter how many people I name I will forget someone, so thank you to all my mentors and peers for the engaging discussions and invaluable feedback.

I would also like to thank all the undergrads I have supervised, without whom I would never have finished so many great projects: Blackfoot Team: Danny and Liyang, Citizen Science: Eric, Lina, and Julie, and BERSt: Maya, Mantaj, Shawn, Flora, Avni, Poorvi, Luna, and Minh.

I would also like to thank my friends and partners throughout my studies. My best friend Elle who is always there to talk me off a ledge, support me when there is a real problem, but also tell me when I am being absolutely ridiculous and unreasonable. She is also always there to tell me to work less, even though she knows I will never listen. Thank you to Correy, who is basically the only reason I know how to code and code well. Everything I know, and everything that makes me a good coder I learned from him. Thank you to Paul, who, in classic French fashion, will blatantly tell me with no sugar coating when my results are bad but will also listen to 100 hours of my voice saying the same sentence over and over again until I get it right. Finally, thank you to my parents, who may not think academia is the best route but are happy to see me succeed either way; whenever I want to give up, I hear my dad's voice in my head (when I failed chemistry my first semester at SFU 15 years ago), "do you really want to give up... you know where people who give up go... to Subway School of Sandwich Making, that's where."

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Imagine you are an immigrant family arriving in a new country. There are pressures to have your children learn the new language and conform to societal norms, but you also want to maintain your cultural heritage and preserve your native language. An important choice that these families make is whether to continue to speak their native language in their home or to begin speaking the language of their new country. Researchers have found that although immigrant families perceive their native languages positively and express a desire for their children to continue speaking their native language, they worry their children will be isolated, struggle with their peers, and fall behind in their education [2]. Moreover, the maintenance of a native language necessitates enormous commitment and encouragement from both parents [3] to both transfer language knowledge and nourish a positive sentiment towards their first language. As such, there is a growing need for support in learning new languages and the maintenance of native languages not only among immigrant communities but also in the case of bilingual countries such as Canada, where a shortage of French language teachers is ubiquitous across the country [4, 5], and in the perseveration of Indigenous languages [6–8].

For several years, the robotics academic and industrial communities have proposed that social robots would be particularly apt for use in language acquisition, as this task is most effectively learned through social interactions [9–11]. There has been rather extensive academic work on the use of social robots for language learning; for a survey, see [10], and for a large-scale project, see L2TOR[1]. Yet, there has been little to no work on improving these robots' voices within robotics and speech communities. Speech is the primary means of social communication among humans. Therefore, it is particularly surprising that little care has been given to designing voices specific to second-language (L2) listeners for teaching tasks. Often out-of-box commercial or on-robot text-to-speech (TTS) [12–18], or recorded human speech is used [19–23] for robot assisted learning. These TTS systems are always assessed by native speakers, and although some work has been done to improve clarity,

[1] http://www.l2tor.eu/

Figure 1.1: Example deployment of our TTS system. The Miroka Robot is using emoji's to add expressivity when reading a story to children in a high energy, noisy environment. At the same time it will adapt it's voice to be heard over the background noise and music.

it is always in the context of background noise [24–28] and not listener comprehension or language ability. To the best of our knowledge, only one previous work has looked at how to adapt a robot's voice to a user's comprehension abilities and ambient noise [29]. Yet, they did not look deeply into mechanisms to support second-language comprehension specifically.

To improve the perception of synthesized voices for second-language problems with TTS voices need to be addressed to aid the comprehension of L2 listeners. The first problem is the technical limitation of deploying voices on robots in the wild; we need voices that are stable, fast, and lightweight; second, we must have an expressive voice that is able to engage users; third, we need the voice to be usable in multiple ambient and social contexts while remaining audible and appropriate; and fourth, we need a voice that is specifically tailored to the comprehension capabilities of L2 listeners.

1.1 Problem 1 - Hardware Limitations

Robotic platforms are frequently limited by hardware and connectivity, making state-of-the-art speech models impractical for deployment. As such, hardware capability continues to be a bottleneck in the explosive field of artificial intelligence (AI), where new,

faster, and more efficient systems are often achieved at the expense of reliability, security, and cost [30]. Many social robots deployed in HRI studies (e.g., Nao and Pepper: Aldebaran, United Robotics Group; iCub: Istituto Italiano di Tecnologia; LOVOT: GROOVE X; Furhat: Furhat Robotics) do not have an on-board GPU or TPU. A new generation of robots with GPUs is becoming available (e.g., Reachy: Pollen robotics; Mirokaï: Enchanted Tools), yet less than 20 (7%) of the 263 robots available on IEEE Robots database[2] specify that they are equipped with GPU hardware. Given that robots often run many systems at once (navigation, balance, vision, grasping, speech, and more), deploying foundation models with an enormous number of parameters on-board is often infeasible. AI models deployed on robots ideally have a small footprint and require a limited amount of resources at inference time [31, 32]; it takes time and careful planning to balance the computational usage of all these models simultaneously. In addition, the ecological footprint of cloud computing [33] and the production of chips needed for powerful GPUs [34] has become increasingly concerning. To reduce costs and the use of power and resource consumption for an ecologically and economically sustainable future in robotics, we must strive to develop smaller and more efficient AI models.

Hardware limitations are often circumvented through the use of cloud computing. Although this may be an adequate solution for in-lab research deployment, in-the-wild, one cannot always rely on a steady nor fast internet connection. This can lead to latency or a complete inability to deploy interactive systems in-the-wild. In addition, using cloud resources can result in several security and privacy concerns [35, 36]. Both subscriptions to cloud services and the inclusion of powerful hardware increase the cost of a robot, increasing the barrier to entry and decreasing the equity of using powerful models in robotics. There is, therefore, a pressing need for lightweight, efficient, and customizable speech models tailored for robotics.

To address this limitation, we introduce the EmojiVoice toolbox. This free, open-source toolkit enables social roboticists to build their own custom TTS with only 3 minutes of training data per voice style. The TTS is small (20.9M parameters with a 78MB checkpoint), efficient, and runs in real-time, even on a CPU.

1.2 Problem 2 - Expressive Voice

It is known within pedagogy that immediacy (i.e., a decrease in the perceived psychological distance between a teacher and their students) and expressivity, both verbally and non-verbally, can increase both affective and cognitive outcomes in the classroom [37–39]. This has been confirmed for virtual and robotic teaching agents as well [40–43]. Yet, in these studies, similar to other studies of expressivity in robot voices outside of the context of

[2] `https://robotsguide.com/`

teaching [44], the range is often limited to changes in intonation and out-of-box voices, such as those built into Nao and Pepper or commercial TTS such as Amazon Polly[3]. These voices may contain, for example, only one expressive "joyful" voice or claim to be "expressive" without providing the user with control over this expressivity. The use of a high-pitched and "joyful" voice[4] is common when attempting to give the impression of an expressive robot [45]. Although it is true that more engaging speakers use a higher-than-average pitch and greater pitch range [46], joy is only a single expression, and humans do not convey expressivity through one single emotion instead, they change their expression over time [47, 48]. Moreover, without any fine-grained control over the expressivity of these voices, tasks requiring extended periods of speech, which may commonly occur in a teaching environment, e.g., telling a story or giving a lesson, will use the same tone of voice throughout the interaction, without considering the appropriateness of this generalized "expressivity" for the phrase nor the task.

The expression and interpretation of emotions is an integral part of human speech [49], and the use of expressive speech not only contributes to a more pleasant interaction but also increases engagement and learning outcomes in a classroom environment. Although "expressive" and "emotional" TTS exist, they have limited controllability, and their use in complex and long-term interactions is poorly understood. Moreover, they are often not customizable, not allowing users to add their own expressive styles relevant to their use cases. There is a need for a synthesized voice that is not only "joyful" but temporally expressive, selecting the correct expressive style given the task and the linguistic context and understanding how and when different expressive styles should be deployed.

To create a voice that remains expressive over time, we built several temporally variable, expressive TTS voices using our EmojiVoice toolbox. In addition, we explored, through case studies, when and where temporally variable expressive and singularly expressive voices are most effectively deployed.

These first two problems are addressed in our in-review article: *EmojiVoice: A TTS Toolkit for Real-time Expressive Speech on Robots, submitted to Robotics and Automation Letters* [50] and a demo: *Take a Look, it's in a Book, a Reading Robot in the Companion Proceedings to IEEE/ACM International Conference on Human-Robot Interaction 2025* [51].

1.3 Problem 3 - Social and Ambiance Adaptation

To ensure a robot voice can be deployed in multiple environments, from a child's bedroom to a noisy classroom, we must understand what adaptations improve the perception of robot

[3]https://aws.amazon.com/polly/

[4]http://doc.aldebaran.com/2-4/naoqi/audio/altexttospeech-tuto.html?highlight=joyful

voices in context. When it comes to humans, from the moment we wake up to the moment we go to sleep, our day contains a variety of social situations in different contexts. For example, our first stop may be to the local bustling café, picking up our usual coffee to start the day. The end of the day may involve a date night at a fancy romantic restaurant or a loud, rhythmic nightclub. In either case, these environments contain different lighting and music to set the mood and people with whom we, hopefully, have the pleasure of interacting.

In all such situations, humans have the unconscious ability to adapt their voice appropriately to the different physical and social characteristics of the environment. A significant portion of adaptive communication results from non-linguistic vocal features that can alter the meaning of phrases [52]. A waiter may ask a customer, "Can I take your order?" and this phrase may carry a different connotation depending on whether the waiter asks the phrase in a casual cafe during a rush or an upscale bar with highbrow clientele. As such, the ambient surrounding plays a role in how one changes their voice to convey meaning to others, and being able to reproduce and interpret these non-linguistic features plays a significant role in human social intelligence.

The need for a robot to adapt itself appropriately in physical and social ambient environments proves crucial when integrating robots into humans' everyday lives [53–55]. We suspect that correct adaptation could improve both the robot's perceived awareness and intelligence. Yet, the exact way to adapt robotic voices has not been thoroughly explored due in part to the difficulty in collecting clean recordings of realistic data in noisy and crowded environments, i.e., to separate voice and background noise. As such, source voice data for TTS is typically recorded and tested in quiet office environments.

To study human adaptation while retaining high-quality, clean audio, we utilized a readily available video-conferencing platform equipped with ambient sounds to aid actors in adapting their voice to fit the target environment while recording from participant's headphones. We then explored how the participants adapted their voices and designed a set of ambient-specific voices to mirror these features, finally conducting a comprehensive user perception study of the voices on a robot in varying virtual environments.

The third problem is addressed in our article: *Read the Room: Adapting a Robot's Voice to Ambient and Social Contexts, published in the Companion Proceedings to the IEEE/RSJ International Conference on Intelligent Robots and Systems 2023* [56], as well as a workshop paper: *Read the Room: Adapting a Robot's Voice to Ambient and Social Contexts, published at Sound for Robotics at the IEEE International Conference on Robotics and Automation 2022*, and a demo: *I'm a Robot Hear Me Speak in the Companion Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction 2023* [57].

1.4 Problem 4 - Second Language TTS

Finally, given the understanding that humans modify their voices to improve interactions, we ask how we can modify TTS to suit the perceptual needs of second-language listeners. The use of voice-based technology, both in perception, production, and end-to-end systems, is seeing expansive growth. These systems can help improve accessibility, such as in-home automation systems for the physically and visually impaired [58, 59], or as a socially supportive device to allow elderly individuals to maintain their autonomy [60]. However, there remains much work to ensure the ethical validity of such systems in terms of transparency, bias, fairness, and many other factors [61]. One aspect of fairness lacking in systems with synthesized speech is their inability to adapt to those with different hearing and comprehension abilities, such as L2 listeners. In human-to-human interaction, the ability of a speaker to adapt to an interlocutor is invaluable. Humans will modify their speech to interlocutors with reduced comprehension abilities, e.g., babies [62] or L2 speakers [63]. Moreover, taking cues from multiple modalities, a speaker is able to perceive when they are not being understood and make adjustments to their speech production to increase clarity both in terms of linguistic [64, 65] and para-linguistic (e.g., emotional [66, 67]) contexts. There is, however, evidence that human adaptation to L2 listeners does not, in fact, aid the L2 listeners in their comprehension [68, 69]. Therefore, to improve the fairness of voice technology and move towards adaptive synthesized speech, we must first understand how speech can be adjusted to facilitate comprehension in a data-driven, perception-based manner.

In English, L2 speakers often struggle to differentiate tense (long) and lax (short) vowels. For example, French and Japanese first language (L1) speakers will often replace the lax /ɪ/ (ship) with the tense /i/ (sheep) [70, 71], while Chinese speakers replace both vowels with the Chinese vowel /ɪ/, which requires a higher and more frontal position of the tongue. Classically, vowel perception is considered as categorizing speech sounds by comparison with a pre-learned auditory representation, presumably a spectral one in formant space [72]. However, it is widely documented that perception also operates relative to its acoustic context [73, 74]. For instance, following a phrase spoken quickly, a sound can be perceived to be longer [75]. Yet, how such context effects can interact or compete, and the exact timescale at which they occur, remains poorly understood [73, 75–77].

To improve the comprehension of TTS from an L2 perspective, we use the data-driven approach of psychophysical reverse correlation [1] to reconstruct the prosodic profile of these difficult vowel sounds over a sentence for both L1 and L2 speakers. Accordingly, we ask: How can the perception of vowels that are known to be difficult for L2 speakers be controlled and comprehension be improved through modifications of pitch and duration across a word or phrase? We then apply these learned profiles to be automatically applied in a TTS system to improve comprehension for L2 Listeners.

This problem has been addressed in our article: *Mmm whatcha say? Uncovering distal and proximal context effects in first and second-language word perception using psychophysical reverse correlation at Interspeech 2024* [78], an article under review: *You Sound a Little Tense: L2 Tailored TTS Using Durational Vowel Properties for Interspeech 2025* [79], and an article in preparation for submission to PNAS.

Taken together, this thesis' contribution is a lightweight TTS with temporally-variant expressivity, social and ambiance adaptation, and L2 clarity mode, providing the building blocks for a synthesized robot voice that can improve human-robot interactions with L2 English speakers.

1.5 Thesis Organization

In the rest of this manuscript, Chapter 2 presents a review of existing literature in the domains of 1) lightweight and expressive TTS, 2) ambient adaptive robot voices, and 3) understanding linguistic features of English vowels for first and second-language speakers. This provides a theoretical background for our work, both providing evidence for the methods used and displaying gaps in the existing literature. Chapter 3 then presents our EmojiVoice toolbox along with our expressive voices and case studies of their deployment (Problems 1 & 2 above). Chapter 4 explores how to adapt a robot's voice to physical and social ambient contexts (Problem 3 above). Chapter 5 presents the linguistic studies that provide the evidence and background for our L2 TTS and the design and testing of the L2 TTS system (Problem 4 above). Finally, in Chapter 6, we present conclusions, future directions, and suggestions on how these improvements can and should be integrated in the future.

Chapter 2

Related Work

2.1 Text-to Speech Synthesis (TTS) for Robotics: Performance and Long Term Expressivity

Robotic platforms for HRI are frequently limited by hardware and connectivity, which severely constrain the deployment of state-of-art TTS systems on robots. In addition, the HRI context creates a stringent need for voice expressivity and adaptiveness, for which speech models developed in other usage scenarii (e.g. vocal assistants) are often not designed. This creates an opportunity for research and innovation, which only a handful of recent work has started to address.

2.1.1 TTS limitations for robotics applications

State-of-the-art TTS models often contain hundreds of millions of parameters (coqui-XTTS: 750M [80], VoiceCraft: 830M [81], Parler-TTS: 878M [82]). For models of this size, a non-negligible part of the robot's storage space and RAM are required for their deployment. Moreover, due to the large computational requirements, these models frequently do not run in real-time, often having a real-time factor (RTF) greater than 1, meaning it takes at least as long as the duration of the audio to generate the speech, disrupting the flow of a conversation. Yet, large models' size and computational requirements are not the only barriers for robotics deployment of state-of-the-art TTS models. Many large models, especially those able to perform zero-shot voice cloning, such as CoquiXTTS [80] and VoiceCraft [81], are auto-regressive. These models are extremely human-like, VoiceCraft having a naturalness mean opinion score (MOS) nearly akin to the original human voice (VoiceCraft: 4.16 ± 0.08, original human speech: 4.29 ± 0.08 on Spotify speech [81]), yet are not readily controllable in a fine-grained manner. For example, it is not possible to control the duration or pitch on a phoneme level as one can with Fastpitch [83, 84] and Fastspeech-based models [85–87]. Yet these controllable models are considered to have low naturalness [88]. Additionally, auto-regressive models often hallucinate, resulting in pure noise or gibberish [89]. Some

models, such as Parler-TTS [82], have reduced the number of hallucinations but still return nonsensical phrases when met with an unknown word or symbol.

Lastly, many of the high-performance TTS systems are either not open source or lack documentation on modifying or fine-tuning the model. Even those that do, once again noting Parler-TTS and its excellent documentation, require large amounts of data to fine-tune to a specific use case. For engaging and robust human-robot interactions, we need a voice to be lightweight, generate in real time, and be free of hallucinations.

2.1.2 Long-term variable expressivity

Many expressive TTS systems are controlled through a series of complex models and hours of expressive training data [90,91], which is contrary to the need for minimalist systems and is not flexible to fast fine-tuning to a specific use case. Moreover, so-called expressive TTS systems mostly focus on increasing pitch and pitch range to create a more “joyful" sounding voice [90, 91], as previous research has found that these modifications drive expressivity in human speech [92,93]. However, there is a lack of work exploring how this type of expressive voice performs over long periods of speech; for example, does listening to a highly expressive TTS for a long time become monotonous? Recent advancements in long-term TTS have improved smoothing over sentences by including contextual information from previous phrases [94], but do not use this to control the expressivity. A recent dataset has been released to help train Chinese expressive storytelling TTS through multiple annotations, including emotional colouring, with the goal of creating a “storytelling voice" [95]. They, however, do not provide an open-source model nor information on the model size, stability, and synthesis time of their baseline. Further, their use of an encoder to control emotion expressivity does not allow the user explicit expressive control, as may be needed for HRI studies. In Chapter 3, we will address these gaps though a lightweight, customizable TTS with temporally variational expressive control.

2.2 Ambient Intelligence for Robotics: How Humans and Machines Adapt Their Voice to Context

Humans' ability to assess and adapt to their physical and social environment is innate and a primary factor in smooth and pleasant interactions. Yet, in human-robot-interactions, we have yet to see extensive work on the adaptation of a robot, particularly their voice, to a given context. This gap provides an opportunity to expand the research on robotic vocal adaptations, improving the perception of a robot's awareness and appropriateness, and the interaction overall.

2.2.1 Context adaptation in human interaction

In human-to-human interaction, it is well documented that interlocutors will both adapt their speech to the context. One important goal of vocal modifications is creating "deliberately clear speech" [96]. These modifications may notably occur in difficult listening environments [97,98], but may also occur to communicate a specific social signal [99]. Modifications are also often listener-specific, as is the case in infant, child [100–103], hearing impaired [104], and machine-directed speech [105]. Vocal modifications are typically produced without conscious effort to elicit a specific auditory feature; rather, they are produced as a result of achieving the aforementioned goals. Moreover, a speaker is able to perceive when they are not being understood and make adjustments to their speech production [64–67].

2.2.2 Contextual Voices in robotics

Although TTS has become an inexpensive and efficient means to create realistic voices for a variety of robotic behaviors [106–108], it is not entirely versatile. For many years, the primary concern for TTS was intelligibility. Because of this, state-of-the-art TTS voices could be mistaken for a human voice, but they still lack the ability to adapt to physical and social contexts. Furthermore, much controllability in stable TTS (i.e., those that do not hallucinate) is constrained by Speech Synthesis Markup Language (SSML). Although the available features have broadened and include loudness, pitch, and rate-of-speech[1], it is not clear whether these features are sufficient for a robot to flexibly and automatically adapt its voice to context.

Several studies have shown that humans show preferences for a robot's voice depending on context and task [53,109–111]. In [53], participants rated the appropriateness of different robot voices in various contexts, including schools, restaurants, homes, and hospitals. They found that, even given the same physical appearance, participants selected varying voices depending on context and concluded that a robot voice created for a specific context is likely not generalizable. Further studies suggest the incorporation of context-based methods such as sociophonetic-inspired design [112], acoustic-prosodic adaptation to match user pitch [113], or incremental adaptation of loudness to the user's distance [114].

While human modifications are well studied, their application to robot voices has been limited. One study has explored the adaptation of a robot's voice to its acoustic environment by adjusting volume based on environmental noise levels [115]. Another used the analysis of the ambient annoyance and user information to modify the robot's voice [29]. Overall, the literature has focused on the adaptation of loudness to ambient environments.

[1] https://cloud.google.com/text-to-speech/docs/ssml

In Chapter 4, we expand on the literature by applying knowledge from studies on human voices, where we understand that contextual appropriateness relies on more than loudness features to address the observed need for more adaptive robot voices.

2.3 Second-Language Speech Perception: Evidence from Linguistics

One critical aspect to improve the fairness of voice technology is to make it adaptive to a diverse set of perceptual needs, perhaps foremost facilitating comprehension for second-language (L2) listeners. While much of linguistic and cognitive science research has explored L2 language perception in human-to-human interaction, currently, there is no existing TTS adaptation strategy for L2 listeners. We review here this literature, and identify a data-driven methodology, reverse correlation, which appears well-suited to develop such a strategy.

2.3.1 Vowel perception in L1 and L2: local and context effects

Classically, vowel perception is viewed as a local (i.e., time-resolute) sensory process that categorizes individual speech sounds by comparing them to a pre-learned auditory representation, presumably spectral in formant space [116]. Formants are the prominent frequencies (F0 being pitch) in a sound that determines a vowel's phonetic quality and are a primary cue for L1 speakers [116]. Speech sounds, however, are highly variable within and across speakers, and it is now widely documented that word perception also operates relative to its acoustic context [73]. For instance, following a context spoken quickly, a sound can be perceived to be longer than it actually is [75]. Similarly, if a speaker's first formant (F1) range is low, a subsequent sound that is ambiguous between /ɪ/ and /ɛ/ may be perceived as having a relatively high F1 and thus subjectively sounds more like /ɛ/ [76, 117]. Such context effects appear to be active in both a forward and backward manner - e.g., they apply whether the word begins or ends the sentence or even if the word is mid-sentence [118]. While it appears difficult to help L2 listeners overcome less spectrally-resolute representations by manipulating the spectral content of the target sound itself (although see [119] for a similar idea for hearing-impaired listeners), the existence of such context effects opens the opportunity to use pitch and duration manipulations within and around the target sound strategically to *bias* the perception of difficult L2 sounds into their correct interpretation.

There is debate, however, about the exact temporal characteristics of such context effects in ecological sentences (i.e., how far the effects are from the contextual changes) and whether the location of the word in the phrase affects which speech cue is involved (e.g., pitch v.s. duration) [73,76]. Previous research has, for instance, suggested that the influence of the preceding speech rate on phoneme distinction may be limited to a temporal window of one or two adjacent phonemes [75]. Yet other work suggests that spectral context ef-

fects result from a form of speakers' vocal tract length normalization (i.e., a listener will inherently attempt to judge the length of a speaker's vocal tract to more clearly interpret their specific pronunciation of formants), which benefits from exposure to long-term spectral cues accumulated over possibly several sentences, or even non-auditorily, from a silent video recording of the speaker [77].

In addition, it is poorly understood whether non-native language processing is able to recruit similar acoustic context mechanisms to those found in native language. For instance, Kang, Johnson, and Finley [120] found that French-L2 speakers failed to use vowel context (i.e., to compensate for coarticulation) when judging fricative sounds when such vowels were unfamiliar. Still, other groups of participants (e.g., English and Tamil, or English and Japanese) behaved identically on different types of sound contrasts [121, 122].

2.3.2 Hypothesis- vs data-driven paradigms in speech perception

Determining the acoustic and temporal characteristics of information intake in sentence perception has always been methodologically challenging for speech perception research. Studies such as [123] have used hypothesis-driven experimental paradigms to establish the causal influence of specific cues on word contrasts by systematically varying their intensity. Yet, these methods are applicable only in a limited scope as assessing the relative perceptual weight of several temporal regions of interest quickly becomes impractical [75].

To document such temporal dynamics, several studies have relied on concurrent eye-tracking in visual search tasks (e.g., printed words or objects corresponding to each word interpretation) and compared the time course of eye fixations to the occurrence of contextual cues [76, 124]. However, this type of study remains relatively uninformative when a specific listener or listener group (e.g., L1 vs L2 listeners [120]) fails to show the effect, as the researcher is left wondering what other possible mechanism may be at play. What is needed is a methodological way to automatically extract, in a data-driven manner, a listener's mental representation of what acoustic profile (i.e., what cue and where) drives a specific vowel contrast one way or another.

Psychophysical reverse correlation [125] is an experimental paradigm used to reconstruct the signals and features that form the mental representation of the decision-making taken by humans when responding to a set of stimuli. The procedure involves a participant observing hundreds of stimuli with random modifications to a target dimension, e.g., facial action units for images [126], or social impressions of a speaker's dominance [1] or confidence [127]. From their responses, one can reconstruct the prosodic profile that maximizes the likelihood of responding to one option or the other (for a review, see [128]). A visualization of the reverse correlation process can be seen in Fig. 2.1.

Update index.html

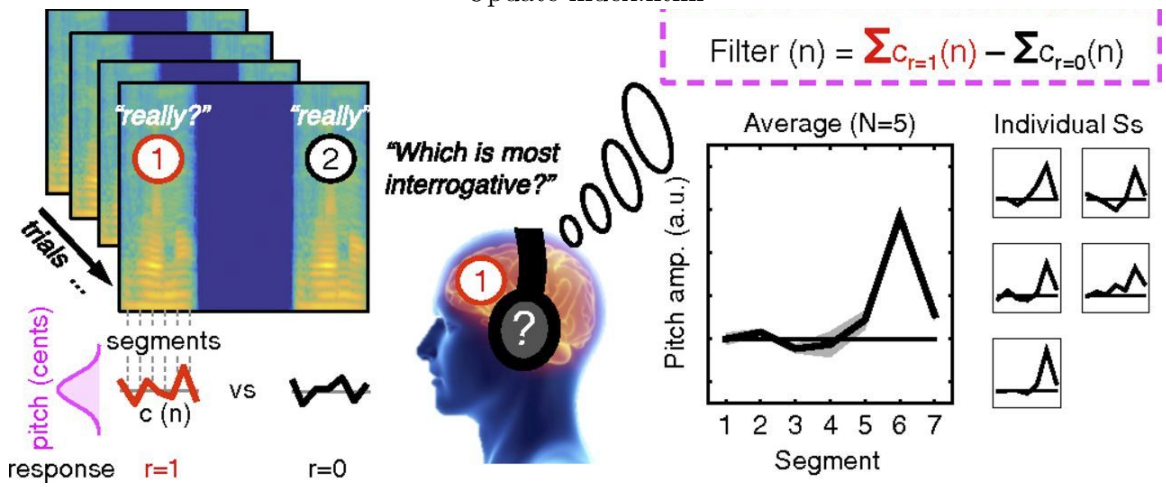

Figure 2.1: Visualization of the reverse correlation process from [1]. (Left) Stimulus with random pitch contours, the participants made a judgment on a pair of sounds (center) resulting in a prosodic mental representations of the judgment.

In Chapter 5, we introduce the use of reverse correlation to investigate whether and how duration and pitch cues, both within a vowel and within the context surrounding a vowel, can facilitate the perception of difficult English tense(long)-lax(short) vowel pairs (e.g., peel vs pill) for L1 and L2 listeners. Finally, we use this novel information to create a "L2 clarity" mode within a state-of-art TTS, and test whether it improves L2 comprehension.

Chapter 3

EmojiVoice Toolbox

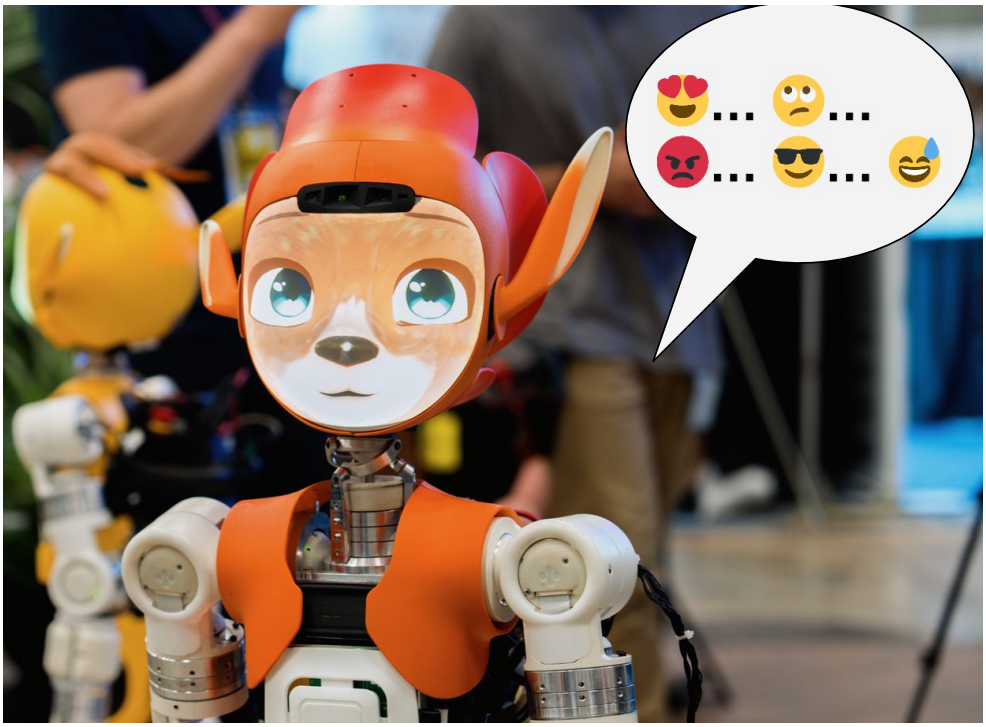

Figure 3.1: Miroka robot speaking with EmojiVoice expressive TTS using emojis to colour the text over phrases.

3.1 Early Exploration of Expressive Teaching Voices

We created a citizen science website called *Now With Feeling*: `https://nowwithfeeling.com/teacherVoice` to get an initial idea of how expressive teaching voices should sound. Here, participants listened to clips of different teaching voices taken from YouTube lectures. They had to rate the voices from most to least engaging. Then, the participants read different text prompts directed at various audiences. For example, one would read a passage from

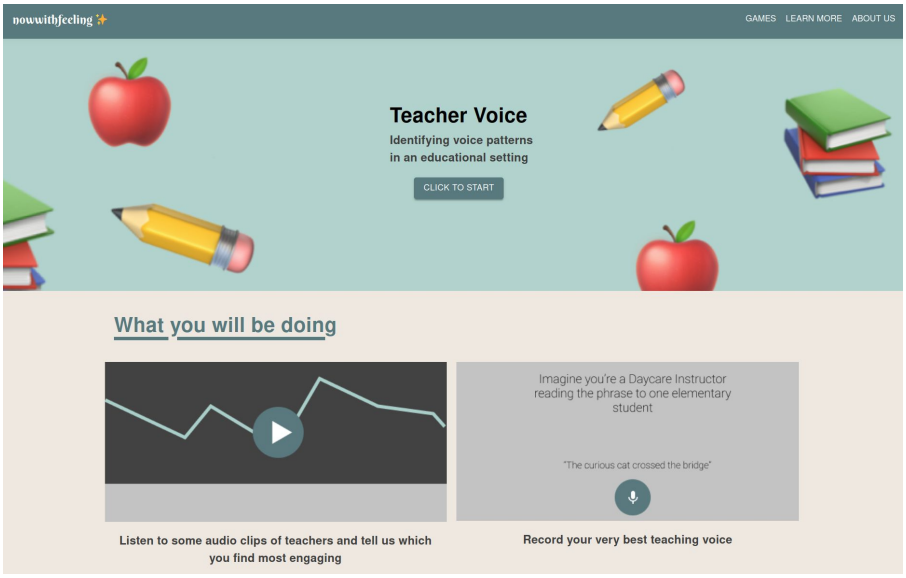

Figure 3.2: Example image from the Now With Feeling webpage where users around the world can submit their best teaching voice.

a storybook to a group of 5-year-olds or read a short lecture passage on mathematics to a university class of 100 students. The participants were prompted with an image of the situation and could practice and play back their speech before submitting the final recording. The site extracted voice features and compared them to voice features of well-known teachers, such as Bill Nye, letting the participants know which teacher they sounded most like. An image of the website can be seen in Fig. 3.2. The website was shared to target teachers, such as on a Reddit page for elementary school teachers.

It was through these audio files that we began to understand the difference between an expressive and temporally expressive voice, especially in a storytelling scenario. We found that the voices that were the most engaging did not read entire pages with a pleasant and excited voice; instead, from phrase to phrase, they changed the type of expression in their voice, keeping us engaged long-term.

To address this and the technical limitations of large TTS models for robot deployment, we built the EmojiVoice toolbox [50]. Our toolbox uses an implementation of Matcha-TTS, a fully open-source, lightweight, and fast TTS model that requires as little as three minutes of data to fine-tune a new voice and has only 20.9M parameters. EmojiVoice runs in real-time on a robot's hardware and can be easily modified to HRI use cases.

Table 3.1: Comparison of open source TTS systems and robot-centric features.

	VoiceCraft	CoquiXTTS	Parler-TTS	FastPitch	Matcha-TTS	EmojiVoice
< 100k parameters	✗	✗	✗	✓	✓	✓
No Hallucinations	✗	✗	✗	✓	✓	✓
Real Time Synthesis	✗	✗	✓	✓	✓	✓
Quick Fine Tuning*	✗	✗	✗	✗	✓	✓
Consistent Speaker Control	✗	✗	✗	✓	✓	✓
Controllably expressive	✗	✗	✓	✓**	✗	✓

*on under 1 hour of fine-tuning data, **control pitch and duration by hand rather than intention or category

3.2 TTS model

We presented, for the first time, a multi-speaker version of Matcha-TTS [88][1] that has been fine-tuned to create EmojiVoice. Pre-fine tuning the multi-speaker TTS was trained on the VCTK corpus [129], with 110 speakers. Behind the scenes, Matcha-TTS is a neural TTS trained using optimal-transport conditional flow matching (OT-CFM). The model is probabilistic rather than auto-regressive, allowing us to circumvent the aforementioned weakness of auto-regressive models (see Section 2.1.1). Moreover, the model is ideal for robot applications as it has a compact memory footprint of only 20.9M parameters, in comparison to 830M for VoiceCraft and 41.2M for Fastspeech 2, and has an RTF of 0.3 (an RTF < 1 is considered real-time [130]) on a Nvidia Jetson AGX Orin 64GB, a GPU with similar power and memory to those applied in production robotics. EmojiVoice takes a label, in this case, an emoji number, along with the input text. This is concatenated both to the text encoder input and the flow-prediction network (decoder) input to contain multiple voices in a single, small (78MB, in our case) checkpoint. The architecture can be seen in Fig. 3.3. Having multiple voices in a single checkpoint reduces space and allows for efficient switching between voice styles without reloading the model. A comparison of Matcha-TTS and other TTS systems can be seen in Table 3.1.

3.3 Emoji Representation

We suggest emojis as a means to prompt expressive styles that can be used for temporal variational control, i.e., selecting a different emoji style for each phrase generated. In 2019, [131] stated that were were 3,019 emojis in Unicode, and this number has now grown to 3,782 in only 5 years. While not all emojis can be considered representations of human expression, considering 106 emojis are smileys alone, with the addition of non-smiley emojis such as “thumbs up" and “flexed biceps,” we have a rich representation of both emotions and social signals larger than, or at least as large as the most commonly used discrete emotion models [132–134]. The variation provided by the standard Unicode emoji set is especially useful in that they can express emotions: 😍, 😡, attitudes: 👍, 😎, mental states:

[1] https://github.com/shivammehta25/Matcha-TTS

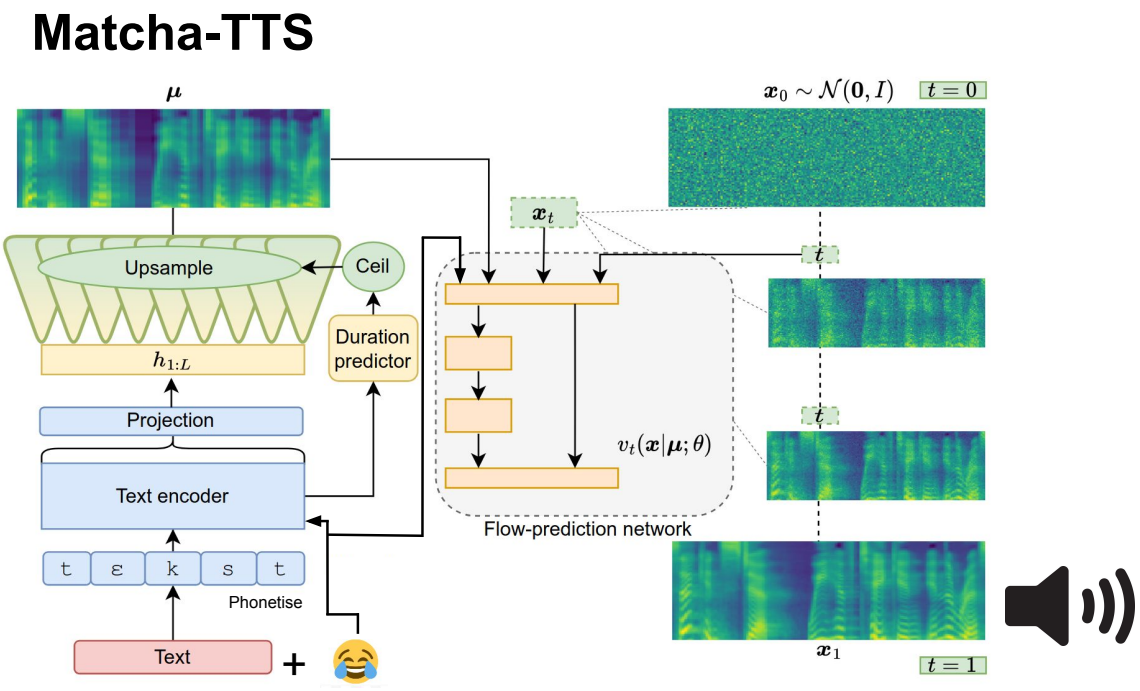

Figure 3.3: Multi-speaker Matcha-TTS with EmojiVoice architecture.

Once upon a time, in the vast Digital Kingdom, there lived a brave Pixel Prince 😎.
His realm was made of glowing grids and circuits, where everything was in perfect order.
Until one day, a fearsome Glitch Dragon appeared, corrupting the code and causing chaos 😡.
The Prince looked up at the flickering skies, worried 🤔.

Figure 3.4: Example of text with emojis. Sample of the script provided to the TTS for Case Study 2: Storytelling task.

🤔, 😕, and bodily states: 🥶, 😴. Moreover, we have become accustomed to the use of emojis in our everyday written expression; for example, almost half of all Instagram posts with text contained an emoji in 2015 [135]. This has resulted in a large amount of the training data for large language models (LLMs) containing emojis along with the text; hence, LLMs can readily produce emojis within text in a rather convincing manner [136,137]. Although LLMs are large and costly, they are increasingly incorporated into robotic conversational systems and are a promising option for generating dialogue [138–141]. This means that emoji-based voice selection can be easily integrated into systems using an LLM with a negligible added cost, where the emotional intention is interpretable by humans. Thus, we used emojis to 1) collect emotional vocal expressions in humans to generate training data for our expressive TTS (Fig. 3.6) and 2) control the TTS voice selection on a phrase-by-phrase basis to ensure long-term temporal variability as seen in Fig. 3.4.

To select the emojis, we aimed to have more positive than negative emojis, unlike the basic emotions, which are primarily negative. We selected an initial list of 23 emojis grouped

Figure 3.5: Emojis used for voice prompting and voice selection.

- The sunlight dances on the lake like diamonds. 😍
- You really think I'm just going to let this slide? 😡
- Seriously? You're going with that excuse again? 🙄
- I thought things were finally getting better! 😭

Figure 3.6: Example prompts for voice actors for data collection and training.

into negative, neutral, and positive. We further filtered the list with a goal of <15 emojis to maintain a reasonable workload for the speaker while maintaining a range of expressions. Through initial exploration, we found that maintaining a consistent production of an emoji was essential for reducing the amount of training data and adequately reproducing the tone of the speaker's expression, i.e., higher variability in the data requires more data to capture these nuances. As such, we kept the emojis where the speaker had a precise mapping of an emoji to a vocal expression and, in turn, was able to produce the most consistent voice. We settled on 11 emojis (Fig. 3.5), a repertoire that outnumbers typical emotional voice datasets (e.g., 9 for IEMOCAP [142], 8 for RAVDESS [143], 8 for MSP-PODCAST [144]). We do not claim that the present set of emojis is the optimal set; rather, it is a starting point for researchers to explore, expand, or even limit the expressivity of robots using this representation, and researchers can use as little as 3 minutes of speech per emoji to customize the TTS. Moreover, emojis could be substituted with an emotion label or text prompts, and we invite researchers to use our toolbox to do so if it suits their needs.

3.3.1 Data collection and training

We provide information on data collection for other researchers to extend or retrain EmojiVoice depending on their needs. To collect training data, we created a prompting script for the speaker. We generated 50 sentences for each emoji using ChatGPT[2] with the prompt: *Please provide 50 short phrases that reflect this emoji: X. The phrases should use all the English phonemes and should not be repetitive, using a variety of words.* These phrases appeared on the screen one at a time with the target emoji appended to the end of the sentence, and the speaker pressed a button each time they were ready to record. Example phrases can be seen in Fig. 3.6. The data was split into 40 sentences of training data and 10 for validation. We fine-tuned off of the multi-speaker VCTK checkpoint available

[2]https://openai.com/

with Matcha-TTS[3]. When fine-tuning 11 of the 110 speakers were overwritten, choosing voices that match the target gender. Other speaker numbers remain available after fine-tuning, however, they tend to be an unstable mix of several speaking voices. We trained on a NVIDIA GeForce RTX 4070 GPU with an Adam optimizer and a learning rate of 1e-4, unchanged from the checkpoint, with a batch size of 20. The model was trained for 85 epochs, approximately 20 minutes. Using our toolbox, we trained three checkpoints with emoji voice data from three actors, two female (Paige and Olivia) and one male (Zach). The toolbox extends Matcha-TTS by including 1) Training files setup: examples, raw data, and 3 checkpoints (with and without optimizers), 2) Additional information on the amount of data needed to fine-tune, 3) Scripts to record data, 4) Wrappers to parse emojis in text to prompt the voices, and 5) a conversational agent. Our toolbox and voices are available free and open source[4].

3.4 Automatic Speech-to-Speech System

In addition to EmojiVoice and the code required to fine-tune and synthesize with it, we include an autonomous, pseudo-speech-to-speech interactive agent chaining automatic speech recognition (ASR), a large language model (LLM) and our TTS with emoji selection to create an autonomous dialogue system. We say "pseudo" as each model runs end to end without linking the embedding spaces between models. The interactive agent runs entirely locally and does not require a connection to the internet. There may remain some latency issues in the LLM on low-power GPUs, but it ran in real-time for our testing on an RTX 4060. We use OpenAI's Whisper *tiny.en* model [145] for speech recognition (39M parameters). To mitigate the complexity of deciding when the user has finished speaking [146], we use a push-to-talk system, recording the voice and producing a transcription. The transcription is the input to the LLM, currently Meta's Llama3.2 [147] (1.23B parameters) implemented via Ollama[5] and LangChain[6] as a chatbot. Although we have used the smallest Llama3 model available, this remains a bottleneck in terms of memory footprint. In the toolbox, the voice designer is able to control the 1) prompt to the LLM to a specific use case and 2) emoji set. This prompt controls the concatenation of the emoji to each response. We extract the appended emoji and use a mapping from the emoji to the associated emoji style numbers, providing this to the TTS to synthesize the appropriate voice. The architecture can be seen in Fig. 3.7

[3] `https://drive.google.com/drive/folders/17C_gYgEHOxI5ZypcfE_k1piKCtyROisJ`

[4] https://github.com/rosielab/emojivoice

[5] https://ollama.com/

[6] https://www.langchain.com/

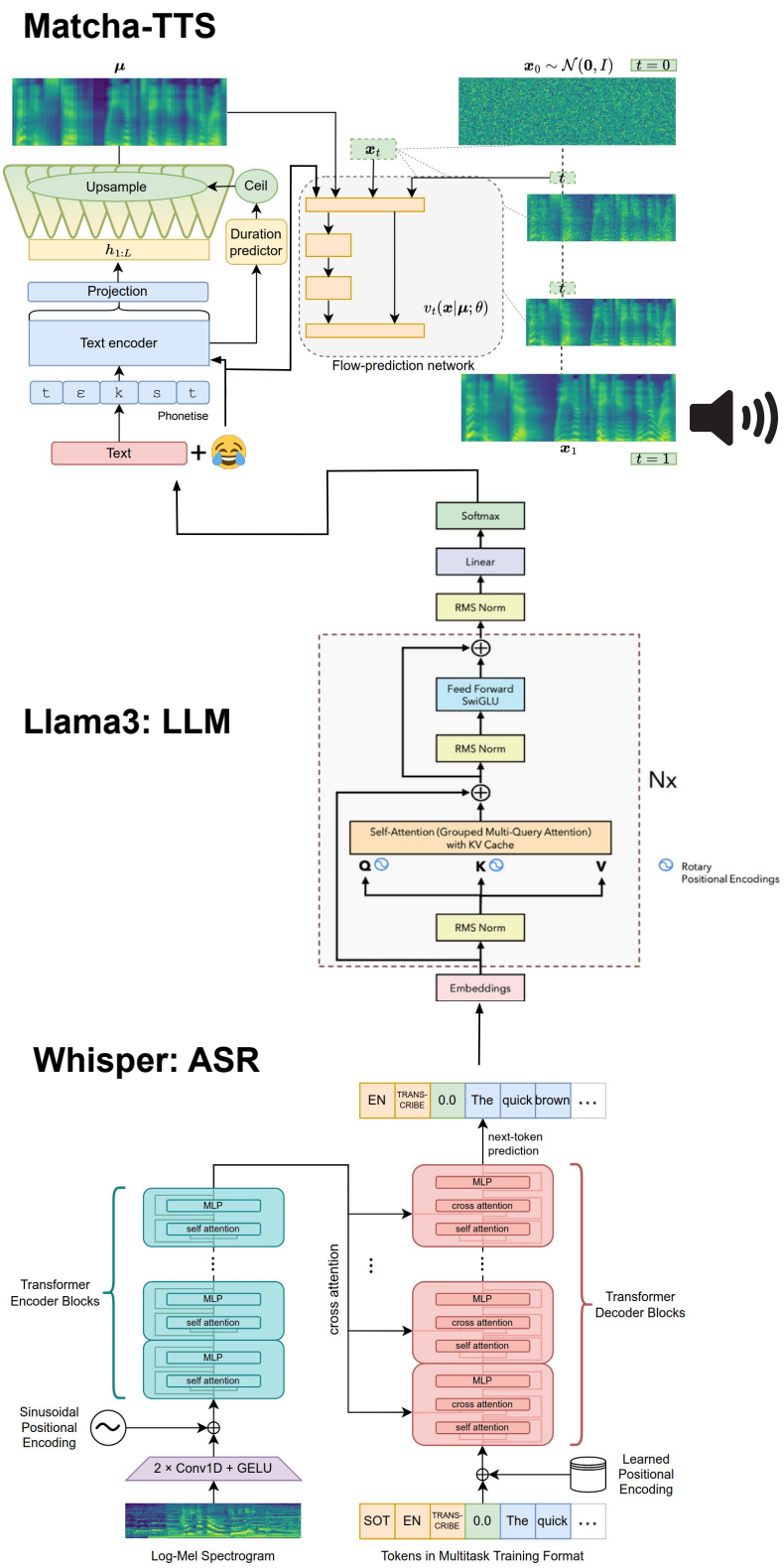

Figure 3.7: Architecture of the pseudo speech-to-speech interactive agent combining: Whisper for ASR (bottom), Llama 3.2 as the LLM (middle), and EmojiVoice as the TTS (top).

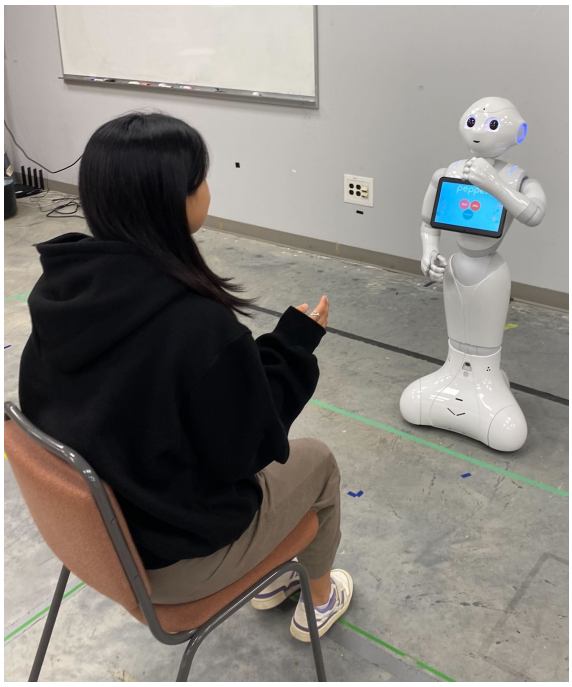

Figure 3.8: Case Study 1 and 2: participant view of the interaction with a Pepper robot.

Figure 3.9: Case Study 3: Participant interacting with the storytelling autonomous speech-to-speech system on a laptop with an image of a Miroka robot.

3.5 EmojiVoice Case Studies

We explore three possible use cases for EmojiVoice to understand how an expressive voice can be effectively deployed in HRI. The first case is a scripted conversation with a robot; the second is a short story told by the robot; and the third is an autonomous interactive agent playing a storytelling game with the user. The first two case studies were conducted on the Pepper and the Miroka (Enchanted Tools) robots, and the third was with an image of a Miroka robot. Each case study compared three voices (using the "Paige" voice): 1. **Baseline**: the original voice from the Matcha-TTS VCTK checkpoint for speaker one for all phrases; 2. **Pleasant**: the EmojiVoice voice that represents the *slightly smiling* emoji, which has an increased pitch range over the baseline as is the focus with other expressive TTS systems, for all phrases; 3. **Emoji**: the complete set of 11 emoji voice styles with each phrase containing an emoji to select a specific voice. Both the Baseline voice and training data for the EmojiVoice were of speakers perceived to be 25-35 years old and female.

3.5.1 Case Study 1: Conversational robot helper

Preparation and setup

For our first case study, participants observed a short, scripted conversation. This case allows us to investigate the perception of short sentences and the interplay with human dialogue.

Alex: Alright, the machine's been sitting for a few minutes, and still no coffee. Pepper, why is it taunting me?
Byte: Wait a second… The brewing chamber! It's not fully locked. Give it a twist until you hear that click. 😮
Alex: Oh, right! Done. Is that it?
Byte: Seriously, Alex? Did you forget to plug it in? The machine needs power, you know. 🙄

Figure 3.10: Case Study 1: Robot Helper with human “Alex", script excerpt.

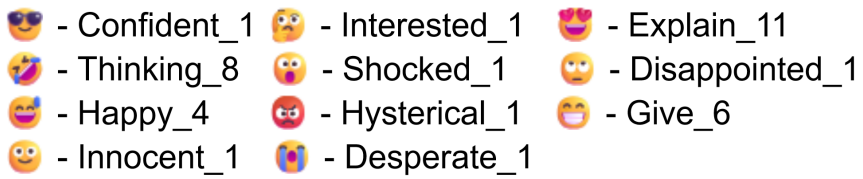

Figure 3.11: Animations played on Pepper in Case Study 1 for each emoji

The script for the conversation was generated with ChatGPT4-o using the prompt: *Can you please generate an interaction between two people, a human and a robot helper, where the robot's lines are always colored with one of these emojis? XXXXXXXXXXX Each of the emojis must be used at least once. It should be about a 5-minute conversation.* A snippet of the script can be seen in Fig. 3.10.

For the Pepper study, the voice was played along with a set of animations[7] (Fig. 3.11). These were chosen to match the emoji's expressivity for each phrase and to be short enough not to cause a lag in the conversation. Basic awareness was turned off to avoid the robot randomly looking around when speaking, which can affect consistency across conditions. For the Miroka experiment, facial animations matching the selected emojis accompanied the voice, and lip sync was on. For both robots, the animations were the same for all voices. The participants knew this conversation was not autonomous, and the researchers controlled both the robot and the sound. The order of the voices was Baseline, Emoji, and Pleasant.

Participants were seated behind the user performing the interaction, as can be seen in Fig. 3.8. Following the interaction with each of the voices, the participants filled out a survey. We used mean opinion scores (MOS) for prosody, intelligibility, and social impression from the MOS-X2 scale [148] and for the expressiveness of intonation as in [149]. Lastly, we questioned the suitability of the voice for the robot following [150]. We used a 10-point Likert scale for all ratings following the MOS-X2 scale to maintain consistency across the questions. Once all three voice interactions had been completed, the participants ranked their preferred voice and explained their choice. Finally, there was an open group discussion to brainstorm and discuss the perception of the voices.

[7]http://doc.aldebaran.com/2-5/naoqi/motion/alanimationplayer-advanced.html#animationplayer-list-behaviors-pepper

Participants

There were 24 participants total (10F, 14M; 16 Pepper, 8 Miroka). The participant demographics can be seen in Table 3.2. All participants completed a verbal consent process as part of ethical requirements, and no personally identifying information was collected. The procedure for all case studies was approved by the Simon Fraser University Research Ethics Board (SFU-REB).

Table 3.2: Participant demographics for Case Study 1 and 2

Category	**Count**
Age Groups	
18-25	4
25-35	16
35-45	4
Ethnicity	
White	10
Chinese	8
West Asian	4
Southeast Asian	1
South Asian	1
First Language	
English	6
Other	18
Daily Language	
English	11
French	5
Chinese	3
Italian	2
Farsi	2
Vietnamese	1
Self-Rated English Proficiency 1(no proficiency) - 5(fluent)	
Score 3	2
Score 4	12
Score 5	10

Results

We used a one-way ANOVA (Type II) followed by a post-hoc Tukey HSD and set $\alpha = 0.05$ to assess differences in the opinion ratings between voices. These results can be found in Table 3.3.

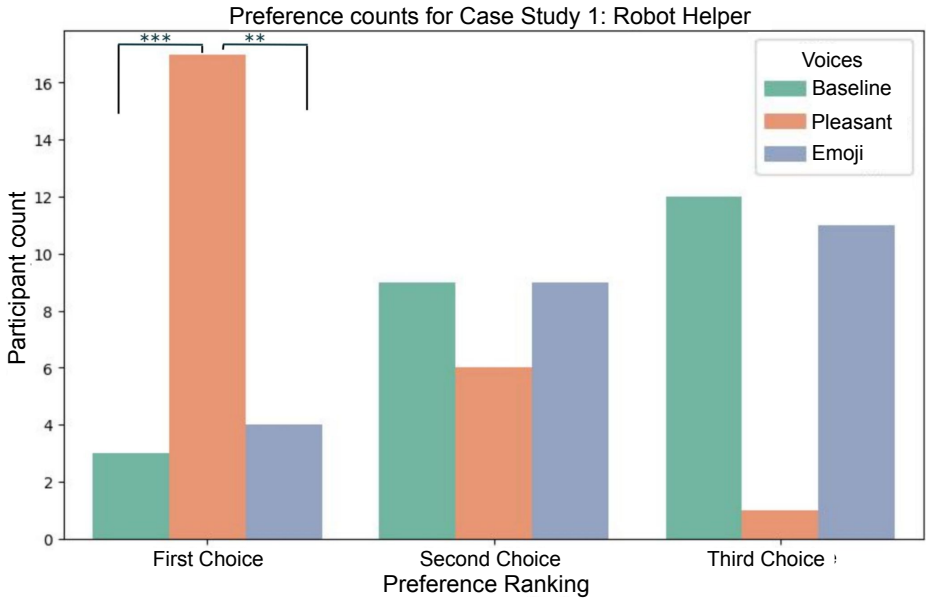

Figure 3.12: Case Study 1: Robot Helper. Participant counts for voice preference of first, second and third choice voices.

In this conversational robot helper scenario, we found that the Pleasant voice was rated as significantly more suitable for the robots (there was no difference between robots) than both the other voices. Both the Emoji and Pleasant voices were found to have a significantly higher social impression than the Baseline voice. Lastly, the Emoji voice was found to have a significantly higher expressivity than both other voices, and Pleasant was more expressive than the Baseline.

We completed Chi Squared Goodness of Fit tests to assess for voice preference, followed by bootstrapped confidence intervals (CI)(Fig. 3.12). We found that the Pleasant voice was preferred as the first choice over the Emoji voice at a 99% CI and over the Baseline voice with over a 99.9% CI.

The discussion of this case study yielded interesting results. Although the Baseline voice is neutral read speech, the participants agreed that the voice sounded "tired and depressed" or "apathetic," and one participant suggested that perhaps a neutral voice would have been better for the study. This same commentary was lacking in further case studies using this voice, where it was indeed perceived as neutral. In addition, the participants noted for the Emoji voice that, "it sounded bit a sassy ... a bit too much in the emotions, like a robot that is judging you." These comments may be a result of the text and forcing the LLM to provide emojis to a robot assistant conversation, resulting in an overly dramatic script. In this case, the participants noted that the pleasantness of the classically "expressive" Pleasant voice made the robot sound "just the right amount of frustrated, like it was annoyed with the task, but it still wanted to keep helping you."

Table 3.3: Case Study 1: Results for difference of mean statistical tests of the MOS Likert ratings.

	Base		Default		Emoji		DF	N	F	P-value		ω^2
	μ	σ	μ	σ	μ	σ				One-Way Anova	Tukey HSD	
xMOS	1.71	1.0	*5.71*	*2.03*	**7.42**	**1.91**	2	24	70.46	< 0.001***	< 0.001*** (E-D, D-B), 0.003** (E-D)	0.67
sMOS	3.58	2.26	*6.63*	*1.81*	*5.96*	*2.22*	2	24	13.81	< 0.001***	< 0.001***(D-B), 0.001**(E-B)	0.26
Suitability	4.54	2.84	**6.96**	**2.33**	5.25	2.09	2	24	6.21	0.003**	0.003**(D-B), 0.04*(D-E)	0.13

1. **Bold** values highlight the voice that is significantly higher than both other voices, *italics* indicate it is significantly higher than one other voice
3. In the Tukey HSD, the results are presented as (higher rated voice - lower rated voice) by the first letter of the voice name
2. Due to space limitations, we only report those voices that were found to have significant differences.

3.5.2 Case Study 2: Storytelling

Preparation and set up

Next, we considered reading a short story. This case allows us to investigate the perception of the voice when stringing together several sentences without interruption. In particular, we suspect that variation in the expressions is an important factor, as speaking continuously with the same "expressive" voice can either become boring or annoying over time. Moreover, the story was generated one sentence at a time to test the real-time nature of the TTS.

The story was generated with ChatGPT-4o using the prompt: *can you please tell me a short story using as many of these emojis as possible XXXXXXXXXXX. The emojis should reflect the emotion in the voice when reading the text and not a physical action. The story should be a fairy tale in a digital realm. I want it to take about 1 or 2 minutes to read out loud.* An excerpt from the script can be seen in Fig. 3.4.

The same setup was employed for the robot. This time, grand gesture animations appropriate for storytelling were chosen for Pepper (Everything_4, Far_3, Thinking_8, ShowSky_11). Facial animations were used similarly for Miroka. The order of the voices was Pleasant, Baseline, and Emoji.

The setup was the same, but this time, the story, without the accompanying emojis, was projected for the participants to read along. This was done to mimic a storybook reading scenario. Following the interaction with each voice, the users filled out the same survey and completed the discussion as in Case Study 1; see Sec. 3.5.1 for further information.

Participants

Case Study 2 used the same participant pool as Case Study 1; see Sec. 3.5.1.

Results

The same assessment was used as in Case Study 1; see Sec. 3.5.1. The results can be found in Table 3.4, and a visualization of the preference counts in Fig. 3.13.

In the storytelling scenario, we found that the Emoji voice was rated as significantly more expressive than the Baseline and Pleasant voices. We also found the Emoji voice to

Table 3.4: Case Study 2: Results for difference of mean statistical tests of the MOS Likert ratings.

	Base		Default		Emoji		DF	N	F	P-value		ω^2
	μ	σ	μ	σ	μ	σ				One-Way Anova	Tukey HSD	
xMOS	2.95	1.57	4.04	2.01	**7.75**	**1.98**	2	24	43.49	< 0.001***	< 0.001***(E-D, E-B)	0.54
sMOS	3.67	2.01	4.96	1.78	*6.29*	*2.26*	2	24	10.07	0.001**	< 0.001***(E-B)	0.20
Suitability	4.17	2.51	5.29	2.42	*6.08*	*2.41*	2	24	3.71	0.030*	0.023*(E-B)	0.13

1. **Bold** values highlight the voice that is significantly higher than both other voices, *italics* indicate it is significantly higher than one other voice
3. In the Tukey HSD, the results are presented as (higher rated voice - lower rated voice) by the first letter of the voice name
2. Due to space limitations, we only report those voices that were found to have significant differences.

have a significantly higher social impression and suitability than the Baseline voice (no significant difference between the robots). The Emoji voice was preferred as the first choice voice over the other two voices at a 99.9% CI.

This time, during open discussion, the opinion was strongly towards the need for a highly expressive voice for a storytelling task. Participants said "you can never be too expressive in storytelling," "the {Emoji voice} is the most expressive, which is very important in storytelling," "I feel that the robot should express emotion according to the story content," and "the {Emoji voice} would be the best for children because of the sudden changes in all aspects of the voice." Additionally, one participant mentioned, "{Pleasant voice} was interesting at first but became boring over time," supporting our hypothesis that over time, a voice that is expressive line by line (in Case Study 2) becomes boring for long text. Yet, there were still some comments on inconsistency in the expression, "{Emoji voice} was entertaining but sometimes the timing, pitch, and emphasis felt off." This may suggest that we need an even lower, more granular level of control over expressivity.

3.5.3 Case Study 3: Autonomous speech-to-speech interactive agent

Preparation and set up

For our final case study, we wanted to better understand how participants view the voices in an autonomous, interactive scenario with the voice. To this end, Case Study 3 was a one-on-one interaction directly between the participant and the system. The exact prompt for the LLM can be found along with the toolbox. The LLM was asked to play a game where it constructs a story by taking turns with the user. The LLM was additionally told to append an emoji that reflects the expression of the phrase to the end of every response it provided. The users played until 2 minutes had passed. As we had a new set of participants interacting one at a time, the order of the voices was selected pseudo-randomly (random while still ensuring all orders were covered) for each participant. As seen in Fig. 3.9, the participants sat in front of a computer screen to complete the interactions. For this more complex, highly divergent interaction where several factors can contribute to participant

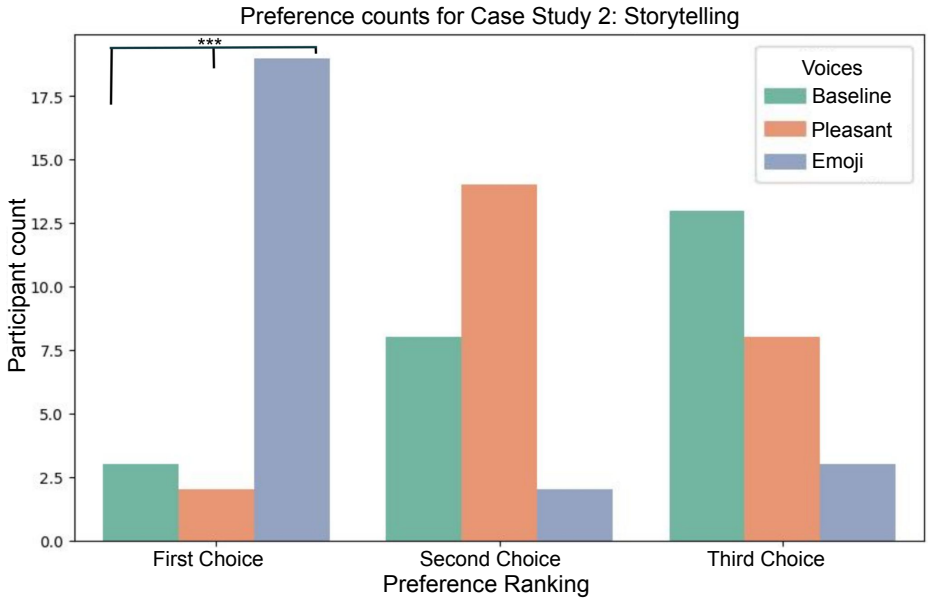

Figure 3.13: Case Study 2: Storytelling. Participant counts for voice preference of first, second and third choice voices.

User: Once upon a time, there lived a princess who loved to eat spaghetti.
Byte: She was very kind and gentle. 😍
User: All of the people that lived in her land loved her.
Byte: They celebrated every day with a big feast, including a big spaghetti dinner. 🤣

Figure 3.14: Case Study 3: Story Building, example interaction between a user and the LLM "Byte".

perceptions, we opted for only a qualitative discussion and assessment of the voices. An example script can be seen in Fig. 3.14.

Participants

There were 8 participants in total (5M, 3F). Participant demographics and be found in Table 3.5. All participants completed a verbal consent process as part of ethical requirements.

Results

As each user interaction was different due to the unique improvisation of the user and system, we opted only to complete an open interview for this case study. From participant discussions, we learned that although the participants were aware that the study's overall goal was to assess the voice, they found their motivations for engagement were highly motivated by the LLM and the setting of the game rather than the voice. One participant said, "It was really fun, but the reason I wanted to keep going was I thought the game was

Table 3.5: Participant demographics for Case Study 3

Category	Count
Age Groups	
18-25	2
25-35	5
35-45	1
Ethnicity	
Arab	3
White	2
South Asian	2
West Asian	1
First Language	
English	1
Other	7
Daily Language	
English	3
French	2
Arabic	2
Chinese	1
Self-Rated English Proficiency 1(no proficiency) - 5(fluent)	
Score 3	1
Score 4	2
Score 5	5

funny; I didn't really consider the voice." Moreover, the performance of the LLM, specifically in its ability to choose emojis, heavily influenced participant perception of expressivity and social impression. One participant said, "It was saying that the princess fell deeper into addiction, but it said it with a laughing voice... it was funny but really inappropriate," giving a negative perception due to a poorly selected emoji. Another noted, "there was an overly sad tone on one sentence (something about piglet for which she was clearly devastated), which created a sense of irony and playfulness," where the emoji was selected appropriately, leading to a more positive impression of the voice. The participants did, however, prefer a more expressive voice: "I expected to hear more expressions when they were talking," speaking of the Baseline voice. Lastly, the participants noted that the system felt like it was in real-time, and any lag seemed appropriate for the robot to be thinking of the following sentence.

3.6 Discussion

When comparing the three cases, we see that EmojiVoice was most appropriate for an expressive, longer-term interaction such as storytelling. This is an interesting result since TTS models tend to avoid using long blocks of text; the longer you hear a TTS, the more monotonous it may sound [151]. Specifically, the expressivity of the Pleasant voice decreased when used in a long-term rather than line-by-line interaction. It does appear that selecting a different emoji for each phrase at run-time lessens this effect. It remains future work to explore the vocal features, such as pitch range, usually associated with expressive speech, in our voices and compare how the performance of just this feature explicitly compares to variability in expressions over time.

Additionally, we found that the perception of a voice can be altered depending on the context of the task and script. Although the same voices were used in the first two case studies, in the first study, the Baseline voice sounded depressed and as if it didn't want to help the user, whereas it just sounded neutral and monotonous when telling a story. The Emoji voice seemed judgmental and sassy in the first study, compared to just being expressive in the second. In the case of an assistant, it appeared that simply choosing a more pleasant voice allowed the participants to feel like the robot was more willing to help with the task, and highly expressive voices are inappropriate, compared to the storytelling where the participants expect a high degree of expressive variability over the phrases to keep their interest.

Moreover, we found it imperative that an LLM, or researcher controlling a WoZ scenario, select the correct emoji or expressivity control to represent a given sentence. Additionally, it was difficult to control the Llama3.2 model only to output our selected 11 emotions, leading too often to the choice of the default voice. Future work could involve specifically training an LLM to a smaller and more lightweight model that has better knowledge of the use case's specific expressions.

Chapter 4

An Ambiance Appropriate Robot Voice

In the previous chapter we built a lightweight TTS with long term temporal expressivity. In parallel we needed to know how to adapt voices to different environments. This will allow a teaching voice to be deployed in varied physical spaces and social situations. Specifically in [56], we asked: How do we maintain appropriateness and give the sense that the robot is aware of its environment, both in terms of physical and social ambiance?

4.1 Understanding Human Ambiance Adaptation

To engage with this question, we first asked: How do humans change their voices depending on the ambiance? The typical first step in designing a voice for a robot is to hire a voice actor to provide hours of voice samples to synthesize a voice from scratch. Another method is to use deep learning and use methods such as voice conversion to change its style. In both cases, a non-trivial amount of data is needed to train a high-quality voice model. Instead, we collected participant data with induced adaptations to ambiances, taking time to understand how humans are adapting to design our robot voices.

4.1.1 Zoom data collection protocol

We proposed a method of virtual data collection using readily available tools that will allow researchers to collect data with no physical human interaction. Although collecting data in an actual restaurant is preferred, collecting in-the-wild high-quality audio free of ambient background noise is challenging. Moreover, targeted ambient environments are often resource-intensive to obtain and control. As such, one of the novelties of the current study was how we overcame the above hurdles by devising a protocol that mimicked naturalistic ambient environments over Zoom[1].

[1] www.zoom.us

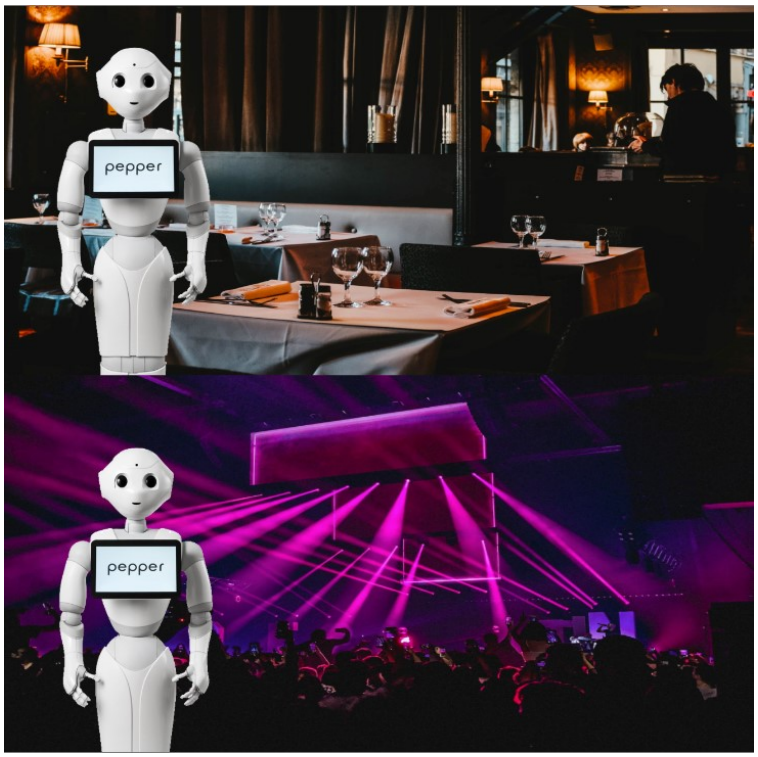

Figure 4.1: Robots may be deployed in varied ambiances, from cozy formal dining to loud nightclubs. How should their voices change?

Zoom is a teleconferencing program that allows individuals to communicate from anywhere in the world. It also allows for the use of virtual backgrounds and sharing of sound. In this study, we asked pairs of English speakers to listen to ambient sounds while conversing with one another in the roles of a waiter and a restaurant-goer using their personal laptop, headphones, and microphone. There were a total of 6 ambient sounds and one additional baseline measure that included no sound and a black background.

In addition to sound, the speakers were asked to change their Zoom background to an image that was pre-selected to match the given ambiance. A brief description of the ambiance (e.g., food available, brightness level, business of location, etc.) and character roles (waiter or restaurant-goer) were provided at the beginning of each ambient condition to induce vivid imagery. There was a one-minute period between each ambient condition to update participants' Zoom backgrounds and prepare for the following condition. This served as a washout to reduce the carry-over effect from the previous condition. The ambient contexts included fine dining, café, lively restaurant, quiet bar, noisy bar, and nightclub. The ambient sounds can be listened to here[2].

[2]https://rosielab.github.io/ambiance

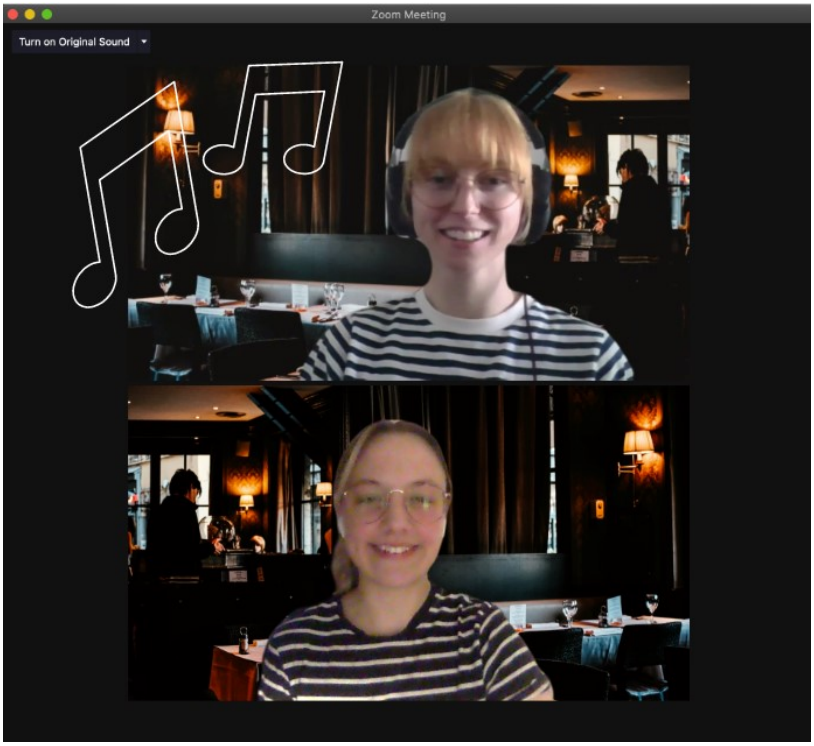

Figure 4.2: Zoom virtual backgrounds and ambient sounds through headphones were used for data collection.

The speakers first read a brief description of their character at the restaurant. They then read from a script that was tailored for the given ambiance, i.e., food and drink choices matched what is usually offered at the given restaurant—consistency amongst scripts allowed for comparison of speech features across each condition. An example script from the "fancy restaurant" can be found in Appendix A. The subtle differences between scripts were solely to create a more realistic environment and reduce redundancy to maintain participant attention. The experiment was repeated in an unscripted manner. However, the unscripted conversations appeared uncomfortable and unnatural and were not included in the analysis.

The dataset consisted of 8 female undergraduate students (age not collected) with experience in improvisation, theatre, or customer service. We used only female voices (biological sex) to match the source voice of Pepper's TTS, which is from a female voice actor. Although the dataset is small, this is often the case in psycho-acoustic studies [152]; we focus instead on the quality of the data through our validation study (see Sec. 4.1.2). Altogether, we collected 837 utterances, resulting in 685 utterances post-validation.

4.1.2 Human voice validation study

The voices were validated using a battery of 7-point Likert scales (1: strongly disagree, 7: strongly agree) to assess the speaker's ability to adapt to the given ambiance as well as the

effectiveness of the virtual ambiance in inducing adaptation. A voice "passage" is a single one-sided conversation of an actor in a given ambiance. Participants proficient in English (N=12, mean age group = 25-35, 4 males, 8 females) validated each passage. Participants listened to each passage overlaid on either: (1) the ambiance the passage was recorded in or (2) a mismatched ambiance. The validation questions loaded onto 4 primary factors: 1. Social appropriateness, 2. Ambient awareness, 3. Comfort, 4. Clarity. The full battery can be found in Appendix B.

Questions on clarity were taken from the well-used MOS [148]. To the best of the authors' knowledge, no validated measures exist targeting the other factors, especially those that can be used out of the context of a larger battery. As we were primarily interested in social and ambiance adaptation, we removed a passage if it received a majority vote of less than 4 (neutral) in any question loading onto social appropriateness and ambiance awareness, meaning that the majority of validators disagreed that this passage was socially appropriate or ambiance aware.

4.1.3 Voice analysis and feature extraction

In order to observe how humans modify their voice, we collected 10 features, which can be grouped by (1) loudness, (2) spectral, including pitch, and (3) rate-of-speech. Our toolbox for vocal feature extraction can be found here[3].

Loudness Features

An increase of vocal intensity, often leading to Lombard speech (an involuntary increase in vocal effort, often due to the presence of background noise [153]), is commonly employed in noisy environments [154]. As such, we collected 3 loudness features: (a) mean intensity, (b) energy, and (c) maximum intensity. Mean intensity and energy features were calculated using the Praat[4] library via Parselmouth[5]. Librosa[6] was used to calculate maximum intensity (power) following the formula provided in Section 1.3.3 of [155].

Spectral Features

We collected 5 spectral features: (a) median pitch, (b) pitch range, (c) shimmer, (d) jitter, and (e) spectral slope. Parselmouth was used to extract (a)-(d). Median pitch and pitch range (the difference between the minimum and maximum pitch in a given segment) are calculated in Hz. Local shimmer and local jitter, variations in the fundamental frequency,

[3]https://github.com/ehughson/voice_toolbox

[4]http://www.praat.org/

[5]https://parselmouth.readthedocs.io/en/stable/

[6]https://librosa.org/doc/main/index.html

Table 4.1: Summary of statistical results of vocal features using an rANOVA.

Voice Feature	DFn	DFd	N	F	P-value
Energy	5	35	8	9.97	< 0.001***
Mean Intensity	5	35	8	15.78	< 0.001***
Max Intensity	5	35	8	12.47	< 0.001***
Median Pitch	5	35	8	2.37	0.06
Pitch Range	5	35	8	0.96	0.46
Shimmer	5	35	8	1.36	0.26
Jitter	5	35	8	1.01	0.42
Spectral slope	5	35	8	9.43	< 0.001***
Speech Rate	5	35	8	1.93	0.12

1. DFn is the Degrees of Freedom of the numerator, and DFd is the Degrees of Freedom in the Denominator for the calculation of the F-statistic.

are perceived as vocal fry and hoarseness, respectively [156]. The spectral slope indicates the slope of the harmonic spectra. For example, a -12dB slope may indicate a falsetto voice and a -3dB slope can indicate richer vocal tones [157]. The spectral slope was calculated using Parselmouth and Librosa with the formula provided in Section 3.3.6 of [158].

Rate-of-Speech Features

We collected 1 rate-of-speech feature: syllables per second. Syllables per second was the number of syllables over the duration extracted using Praat scripts[7].

Trends in Voice Data

The data for this study was collected as a repeated measures experiment. Six treatments and a baseline, each of the ambiances, were applied to each of the study participants. Each pair of study participants was independent; however, within the pair of waiter and customer, we do not have independence, as synchrony and mimicking are expected to occur. We completed repeated measures ANOVA (rANOVA) for each of the extracted voice features. We saw several features that warranted further investigation. Energy ($p < 0.001$), spectral slope ($p < 0.001$) , max ($p < 0.001$) and mean ($p < 0.001$) intensity, pause rate ($p = 0.002$) and mean pitch ($p = 0.06$) were all significant. The results can be seen in Table 4.1.

Radar plots constructed from our collected dataset are displayed in Fig. 4.3. We can observe that although voices in the quiet ambiances appear to share vocal features, those in each of the loud ambiances appear quite different. The average voice in the bright, high arousal lively restaurant with fast music (140 BPM) showed a higher pitch range and energy with a low spectral slope when compared to the other ambiances. The average nightclub voice, by comparison, showed high shimmer, jitter, and spectral slope, with a high pitch

[7]https://github.com/drfeinberg/PraatScripts/blob/master/syllable_nuclei.py

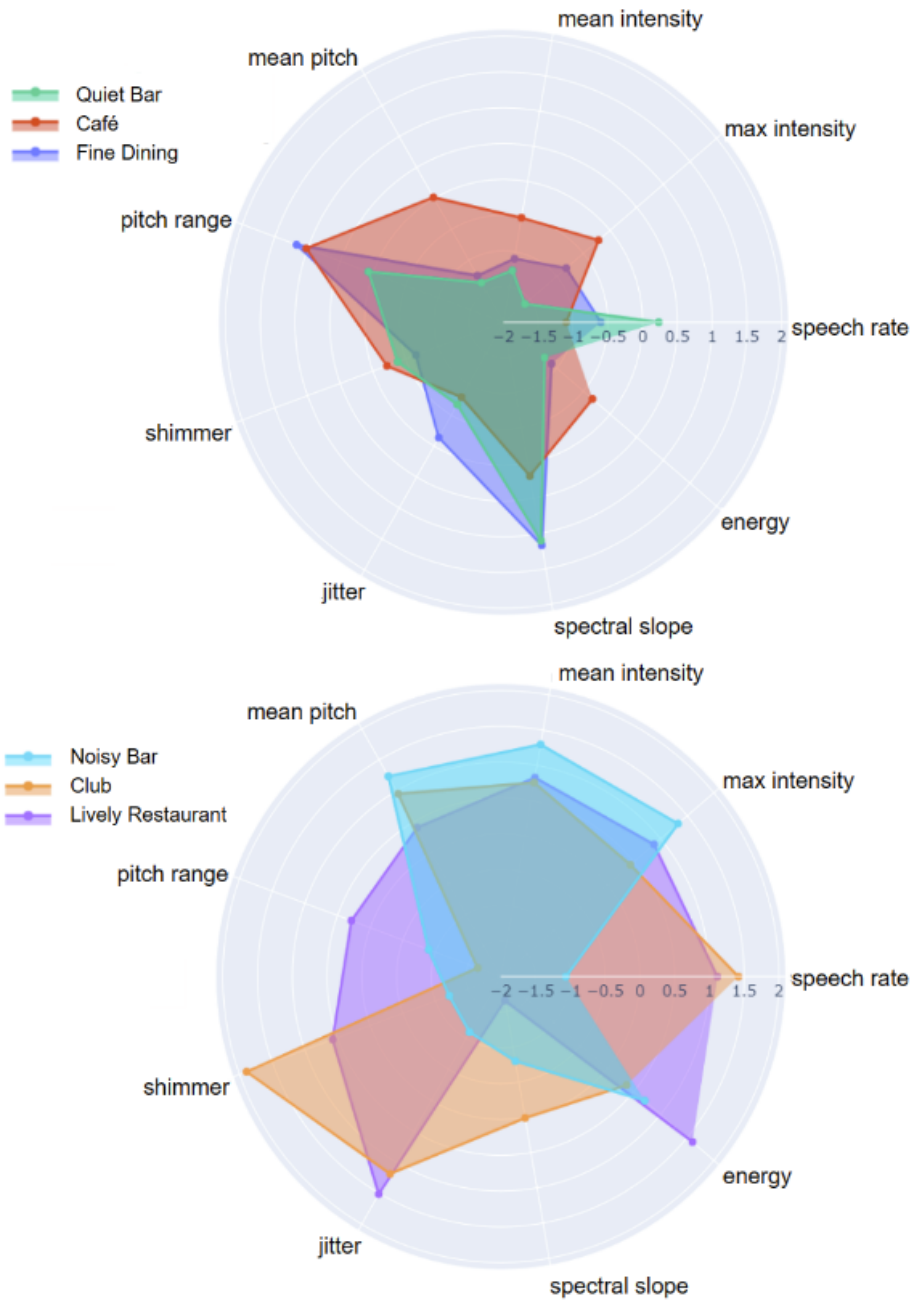

Figure 4.3: Vocal features averaged across speakers as a difference from the baseline. Top: quiet ambiances, bottom: noisy ambiances.

and low pitch range, suggesting a voice consistent with Lombard speech. Lastly, voices in the loud bar showed moderate values for most features aside from high loudness. First, these trends suggest that adapted voices indeed vary by features other than loudness. As loudness is already well represented in the literature, continuing forward, we will freeze loudness features across voice styles to focus on the contribution of other vocal features. Second, we observed that the quiet voices may be a single style.

Given our analysis of human voices, we deemed it necessary to explore further how to adapt a robot's voice to different social and ambient contexts. This need is essential as it can impact how an individual perceives a robot as acceptable in different contexts.

4.2 Pilot: TTS Adaptations

We first piloted adaptive TTS, reflecting the vocal features from our dataset. Our generated voices for the perception study contained 6 TTS voices. These samples were generated using Google TTS[8], which allows for the use of SSML to alter the following elements: (a) loudness, (b) rate-of-speech, and (c) pitch. The first sample was a baseline TTS, which has no alterations of the stated SSML elements (TTS-bl). The next sample was a set of 6 TTS voices that were generated by setting the SSML elements to the average features of all female speakers for each ambiance (TTS-avg). Finally, two TTS samples (TTS-low and TTS-high) were generated with matching loudness and rate-of-speech elements of TTS-avg but differing pitch levels. TTS-low had the pitch element set to one specific speaker's pitch, a female undergraduate student from our dataset (714). TTS-low's pitch sat between the pitch of TTS-bl and TTS-avg. TTS-high's pitch element was set to $pitch(\text{TTS-avg}) + (pitch(\text{TTS-avg}) - pitch(\text{TTS-low}))$.

The perception study leveraged Mechanical Turk and Survey Monkey with 25 Canadian participants who were fluent English speakers, with 100 Human Intelligence Tasks (HITs) completed with a 98% acceptance rate. There were 2 research questions we investigated:

- **RQ1**: How do generated voices compare to human voices within ambiances?
- **RQ2**: How does a data-driven pitch manipulation for TTS impact human perception?

Listeners were first asked to use headphones and calibrate their audio. Participants were told that Pepper was here to take their order and were asked to read one of the 6 ambiance descriptions for a restaurant-goer. After listening to Peppers' voice over the background sound, participants were asked to respond to 2 statements using a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). The following statements were provided: (1) Pepper's voice sounds socially appropriate for the scene, (2) Pepper makes me feel comfortable.

[8]https://cloud.google.com/text-to-speech/

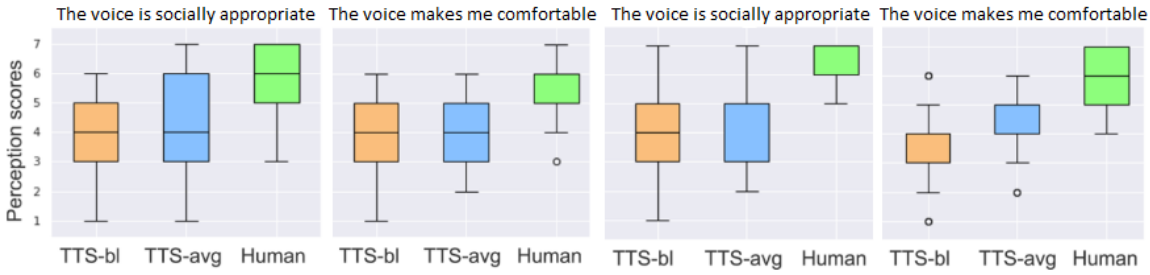

Figure 4.4: RQ 1: Comparison of the perceptual rates of three voice types for fine dining (left two) and night club (right two)

4.2.1 Pilot results

RQ 1 : How do generated voices compare to human voices within ambiances?

TTS-avg, TTS-bl, and a human voice were rated by participants for the fine dining and nightclub ambiances. These ambiances were chosen due to their polarity in formality and loudness. TTS-bl rated the lowest for both statements. Additionally, all generated voices were ranked noticeably lower than the human voice (see Fig. 4.4). This indicated that a human voice, confirmed with a post-hoc Tukey test, was more comforting and socially appropriate for the ambiances, $p<0.05$. Although human perceivers preferred the human voice, they also preferred a TTS that had the pitch altered to the context, compared to that of the TTS-bl. This result supports our goal of choosing a TTS that adapts to the context. This is further supported in RQ 2.

RQ 2 : How does a data-driven pitch manipulation for TTS impact human perception?

TTS-bl, TTS-avg, TTS-low, and TTS-high were overlaid on the café background sound and compared. The TTS-bl was rated the lowest for appropriateness and comfort. The results (as shown in Fig. 4.5) indicate that the human-pitched TTS (TTS-low) was deemed more socially and contextually appropriate, as well as comforting. Altogether, using a data-driven method to alter pitch, demonstrates the result that humans prefer when the pitch matches the current social context and ambient environment.

The pilot results continued to suggest that voices adapted to the ambiance improve the perception of these voices. As such, we continued to a full, on robot study.

4.3 Robot Voice Styles Using Clustering

The TTS voices used in the pilot appeared not to be the best match for a robotic embodiment. Moreover, as the development of a robot voice is resource-intensive, we aimed to find

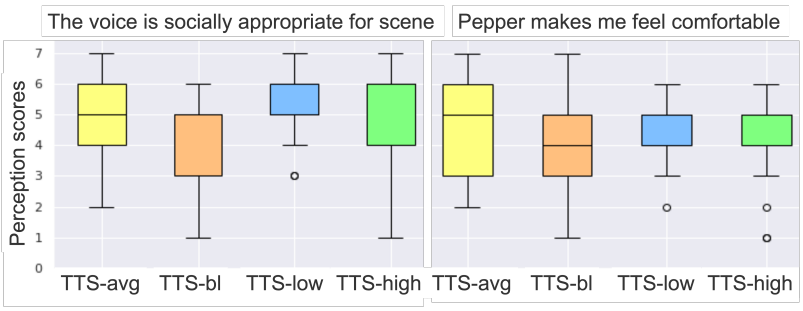

Figure 4.5: RQ 2: A comparison of perceptual rate in four voice types

a minimal set of voice styles needed to work well in our ambiances. We used scikit-learn[9] to scale and cluster our voice data. First, K-means clustering was performed on voice features collected in Section 4.1.1. Robust scalar[10] was used to normalize each feature. Once scaled, the number of clusters, set from 2 to 10, was tested, and a silhouette score was collected to determine the appropriate amount of clusters. Although the silhouette score was lower for a larger number of clusters, when observing the cluster composition, we observed that, more often than not, the number of utterances for each ambiance was equal across clusters. This suggests vocal features for these clusters differ on an utterance level rather than ambiance level. Exploring utterance-level information, such as positive or negative phrases, is left as future work. As such, we determined three clusters was appropriate and that three voice styles would need to be selected or designed.

4.3.1 Human voice cluster analysis

In which ambiances are the voice styles (derived from clusters) primarily used, and what do they sound like? Towards answering these questions, we created two types of plots: (1) for each cluster, the proportion of each ambiance's utterances using this voice style (Fig. 4.6), and (2) a radar plot of each cluster's centers to depict the voice style's characteristics (Fig. 4.7). In Fig. 4.6, we see that utterances belonging to the first cluster occur mostly in the fine dining and quiet bar ambiances and do not occur in our two highest arousal ambiances. We therefore designate it as "calm". The perception rating of each ambiance can be seen in our perception study Fig. 4.9. Utterances in the second cluster occur most often in the lively restaurant, a high arousal, positive ambiance as well as the brightest. We designate this second cluster as "exciting or bright". The third cluster is occurring primarily in the loudest ambiance, the nightclub. We suspect this is a Lombard voice and designate the name "loud". One ambient context of note is the loud bar, which appears to occur almost equally across clusters. This may have occurred for several reasons, including a split in adaptation

[9]https://scikit-learn.org/stable/

[10]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

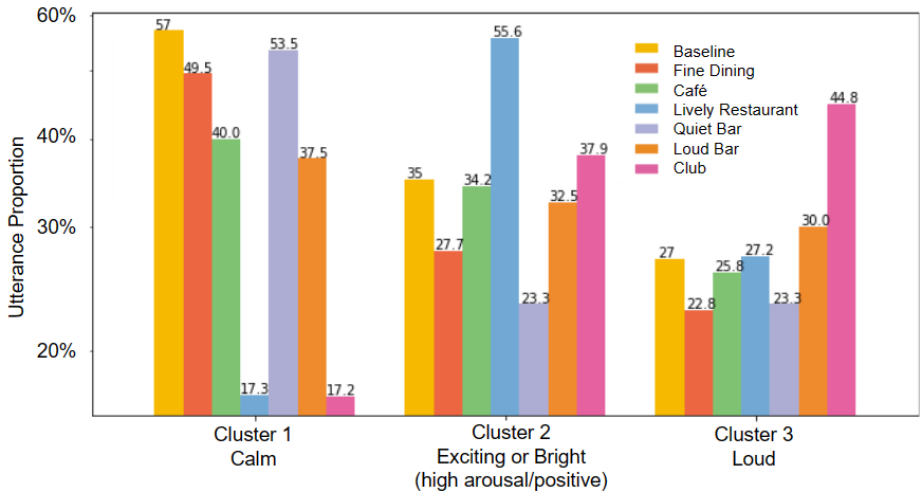

Figure 4.6: Which voice styles are associated with which ambiance? We visualize the proportion of each ambiance's utterances represented by each voice style.

between the server and the customer, different interpretations of the atmosphere by the speakers, or strong utterance level differences for this ambient context.

The radar plot of the features for each cluster's centers (Fig. 4.7) shows that the "exciting or bright" cluster has high energy, speech rate, pitch range, jitter, and shimmer. This appears to fit the assessment of this being a happy and excited voice. For the calm cluster, we see moderate values on most spectral and pitch features with the highest spectral slope, which can indicate a richness in the voice [157]. Lastly, looking at the loud cluster, it does indeed appear to be Lombard: we see the low speech rate, lower pitch range, and high pitch that is quintessential of this voice style. Lastly, we see a parallel from the initial analysis of the human voices (Fig. 4.3), where the quiet voices belong to a single cluster, yet voices in the loud ambient contexts differ based on a number of factors. Our next step is to select robot voices in these styles towards a robot voice perception study.

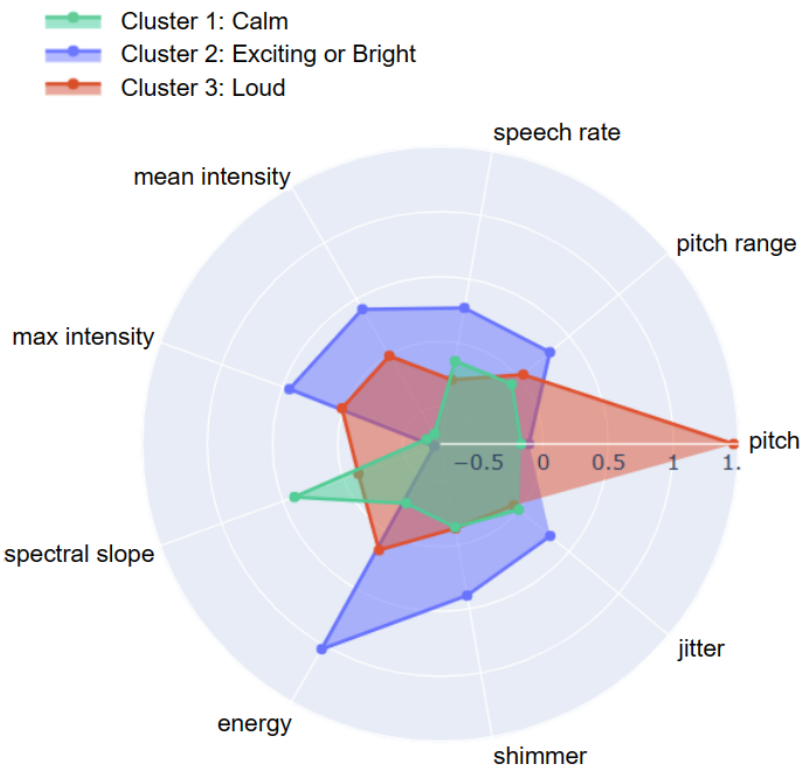

Figure 4.7: What do the voice styles sound like? Feature analysis of the three voice styles derived from human voice cluster centers.

Table 4.2: Comparison of features for voice clusters and Pepper voices.

Human Voice Clusters			
	Pitch(Hz)	Pitch Range(Hz)	Speech Rate(syll/s)
Calm	Medium	Medium	Medium
Exc/bright	Medium	High	High
Loud	V. High	Medium	Low
Pepper Voices			
Neutral	Low(350)	Medium(575)	Medium(4.18)
Joyful	V. High(408)	V. High(699)	High(4.41)
Didactic	Medium(367)	Low(558)	Low(3.71)
"Lombard"	V. High(435)	Low(467)	Low(3.71)

Note: For human voices, the values were assigned based on the value of the cluster center following robust scaling. Values between -0.25 and 0.25 were assigned medium, those over 0.25 high and below -0.25 low. For Pepper's voices, features were scaled separately from the human features using MaxAbsScaler with the values in the table.

Voice Clusters to Robot Voice Styles

We extracted voice features, once again using our voice toolbox, from Pepper's three available voice styles: Neutral, Joyful, and Didactic. Given the human clusters, our next step is to map our voice styles to our robot's voice. We aim to make as little changes to Pepper's voices as possible to demonstrate that, although these voices have been expertly curated to match this robot, they may not be applicable in all situations, and care needs to go into understanding in which context each voice should be applied. We also leave synthesizing new voices as future work, as the proposed data-driven voice selection method does not need access to the robot's original voice actor or a large amount of training data. Therefore, it is more applicable to researchers working in low-resource environments. Moreover, it is important to maintain the identity and overall features that make these voices appropriate for this robot.

A comparison of the cluster features with the features of each of Pepper's voices can be found in Table 4.2. We only include those features that can be modified through Pepper's TTS mark-up language. Again, we chose not to include loudness features as we aim to maintain Pepper's voice at a constant volume to explore how the alteration of non-loudness features can affect the perception of the voice. Based on the results in the table, we decided to equate the "exciting or bright" cluster with Pepper's Joyful voice and the "calm" cluster with Pepper's Neutral voice. For the "loud" cluster, we did not have a particularly close match given the high pitch and low pitch range. We noted that Pepper's Didactic voice had the lowest pitch range and speech rate. Therefore, we decided to build our "Lombard" voice with this Didactic base, increasing the pitch by 130%.

Table 4.3: Lighting (Lux) and sound (Db) settings for each condition.

Ambiance	Formal Restaurant	Café	Lively Restaurant	Quiet Bar	Noisy Bar	Night Club
Lux Level	44	17	17	44	44	0
Db Level	50	50	55	50	55	60

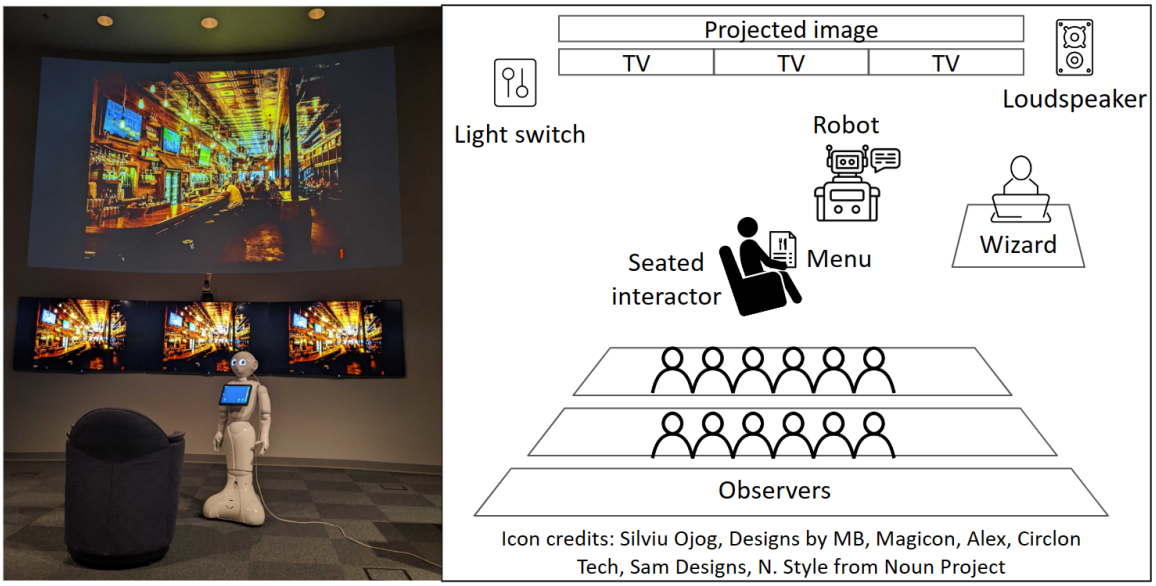

Figure 4.8: Experiment setup for user perception study, and the "noisy bar" condition (left).

4.4 Perception Study

We conducted a perception study to better understand human perception of our selected ambient robot voices. We aimed to address three research questions (RQs):

RQ 1: What voice styles are preferred for the target ambiances?

RQ 2: What ambiance-specific voice styles can increase perceived appropriateness and user comfort?

RQ 3: Does improved appropriateness and comfort occur with an increased perception of competency, awareness, and human-likeness, i.e. overall intelligence?

We recruited 120 students to participate: 71 men, 44 women, 4 non-binary, and 2 who preferred not to answer. The participants were primarily between the ages of 18 and 24 (N=114). Participants were required to have normal hearing, be over 18, and have a proficient grasp of English in order to participate. The students were primarily recruited through an introductory computing science course. All participants provided their informed consent and the procedure was approved by the SFU-REB.

We ran our study in a presentation theatre within a university campus as a controlled proxy for actual restaurants. In order to replicate our desired ambiance, we projected the same restaurant image used to collect voice data on the wall, as well as three TV screens at

the front of the hall. Additionally, we used the ambient speakers available within the room to play the music used in the data collection. Fine dining, café, and quiet bar were set to 50 dB, lively restaurant and noisy bar were set to 55 dB, and nightclub was set to 60 dB using a decibel meter. Lastly, we controlled the lighting by keeping the blinds closed and turning on and off lights to replicate 'dark,' 'warm and dim,' or 'bright' environments. The level of light and volume of the sound used for each ambiance can be found in Table 4.3 and an image of the setup in Fig. 4.8. The ambiances were presented in random order for each trial.

The participants interacted one at a time, for a single interaction with the Pepper robot, while the others observed. For each ambiance, three interactions took place using the script below, each with a different voice style. Pepper was controlled using Wizard of Oz, with Pepper's ALTextToSpeech module. Pepper said, *"Hello, I hope you are doing well. I hope you have had a chance to look at the menu. I recommend the daily special. Anyways, what can I get you?"* The interacting participant (hereafter, "interactor") then verbally replied, choosing an item from the paper menu in front of them. Pepper then replied, *"Sure, I can definitely do that. I will be right back with your order."* During the conversations, Pepper's tablet displayed a black screen, and no gestures were used. Basic Awareness and idle body motions were activated to allow the robot to appear lifelike while also allowing the participants to focus on the voice.

Once the conversation was completed, all participants, both the interactor and the observers, completed an online questionnaire. A black screen was shown during this time, and no music was played. The questions began by asking whether you were the one who interacted with the robot. The remaining questions used a 7-point Likert scale (1: strongly disagree, 7: strongly agree) as follows:

- Pepper's voice is socially appropriate for the scene
- Pepper is aware of the surrounding ambiance
- Pepper makes me feel comfortable
- Pepper sounds human-like
- Pepper is a competent server

To the best of the authors' knowledge, no validated measures exist clearly targeting all of these factors, specifically social appropriateness within a scene and ambiance awareness, especially without the inclusion of numerous extraneous questions. At the end of each ambiance, the participants answered a "scene change" question set while the next ambiance was set up. These questions included which of the three voices they would prefer as their server and questions to assess the accuracy of the ambiance. First, we asked whether "The ambiance matched the prompt" with a 100-point slider from disagree to agree. The subsequent questions were on a sliding scale from 1 to 100 with the prompt "The ambiance is:"

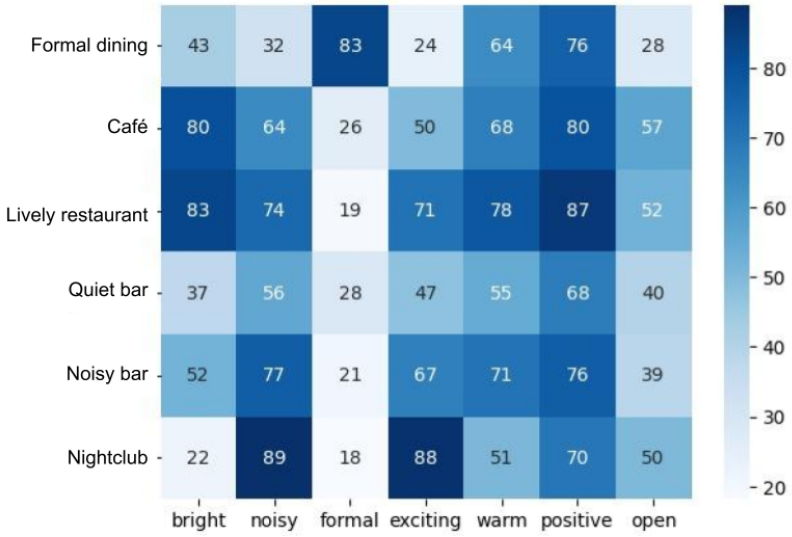

Figure 4.9: Perception of ambient characteristics (rated from 1-100) for each condition based on participant results.

with slider scales of Dark-Bright, Quiet-Noisy, Casual-Formal, Calm-Exciting, Cold-Warm, Negative-Positive, Enclosed area-Open area. We completed 18 of the previously described short conversations for each trial—the three voices from Sec. 4.3.1, Neutral, Joyful, and "Lombard" were used in each of the 6 different ambiances: a fine dining restaurant, a café, a lively restaurant, a quiet bar, a loud bar, and a nightclub.

4.5 Results

We first tested all results for significant differences between the ratings of the observer and the interactor, and none were found; therefore, moving forward, observers and interactors will be treated as members of the same group.

4.5.1 Ambiance analysis

We found that, in general, our physical ambiances were a good match to the given prompt (N=714, μ=81.25, σ=20.10) when rated on a scale from 1 to 100. The agreement was lowest for the quiet bar (N=120, μ=77.34, σ=23.56) and highest for the fine dining restaurant (N=120, μ=85.63, σ=17.86). The average ratings for each ambiance feature on a scale from 1-100 can be seen in Fig. 4.9.

4.5.2 RQ 1: Preferred voices for ambiances

We completed Chi Squared Goodness of Fit tests, followed by confidence intervals (CIs) to assess preference for specific voices in a given ambiance. At $\alpha = 0.001$, we found that a Neutral voice was least often selected as the preferred server in a lively restaurant and a

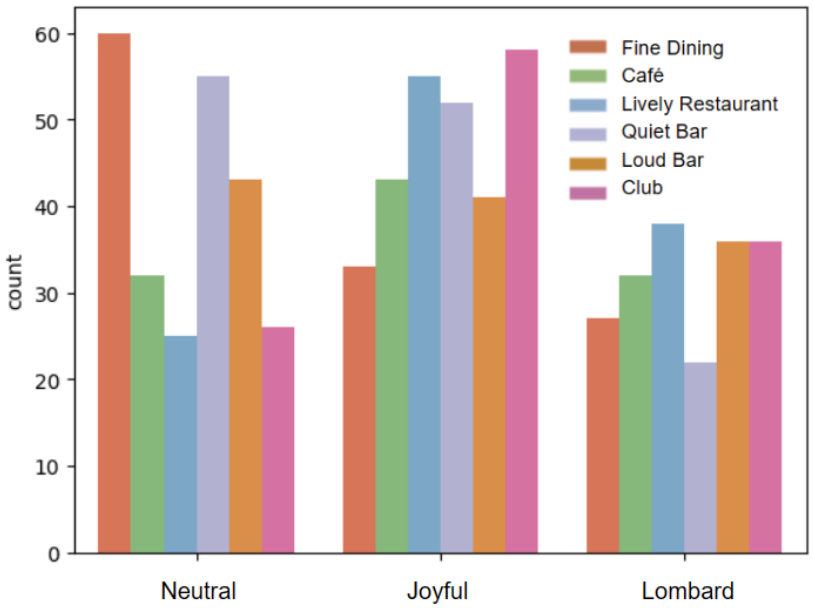

Figure 4.10: Number of participants choosing each voice as first preference for each ambiance.

nightclub and that the "Lombard" voice was least often selected as the preferred server in the fine dining and quiet bar. Moreover, the Joyful voice was least often selected as the last choice in the lively restaurant. So, although Joyful showed no significance as the preferred voice, we know that it is not the least preferred. Full significant results can be seen in Table 4.4, and a plot of first-choice voices can be seen in Fig. 4.10.

4.5.3 RQ 2: Social appropriateness and comfort

For both RQ1 and RQ2, we used a one-way ANOVA followed by a post-hoc Tukey HSD and set α=0.001 to assess improvements in Likert ratings between voices within each ambiance for our targeted questions. A full table of these results can be found in Table 4.5.

We found that a Neutral voice was more socially appropriate for fine dining and that a Joyful voice was more socially appropriate in a lively restaurant, noisy bar, and nightclub. In addition, a Joyful voice made participants feel more comfortable in a nightclub.

4.5.4 RQ 3: Awareness, competency, and human-likeness

We found that a Neutral voice was perceived as more ambiance-aware for fine dining and that a Joyful voice was perceived as more ambiance-aware in a lively restaurant, noisy bar, and nightclub. In addition, the robot with a Joyful voice was perceived as more human-like and competent in a nightclub and more competent in a noisy bar.

Correlation results can be seen in Fig. 4.11. We find the strongest positive correlations between social appropriateness and ambiance awareness in all ambient contexts. Strong

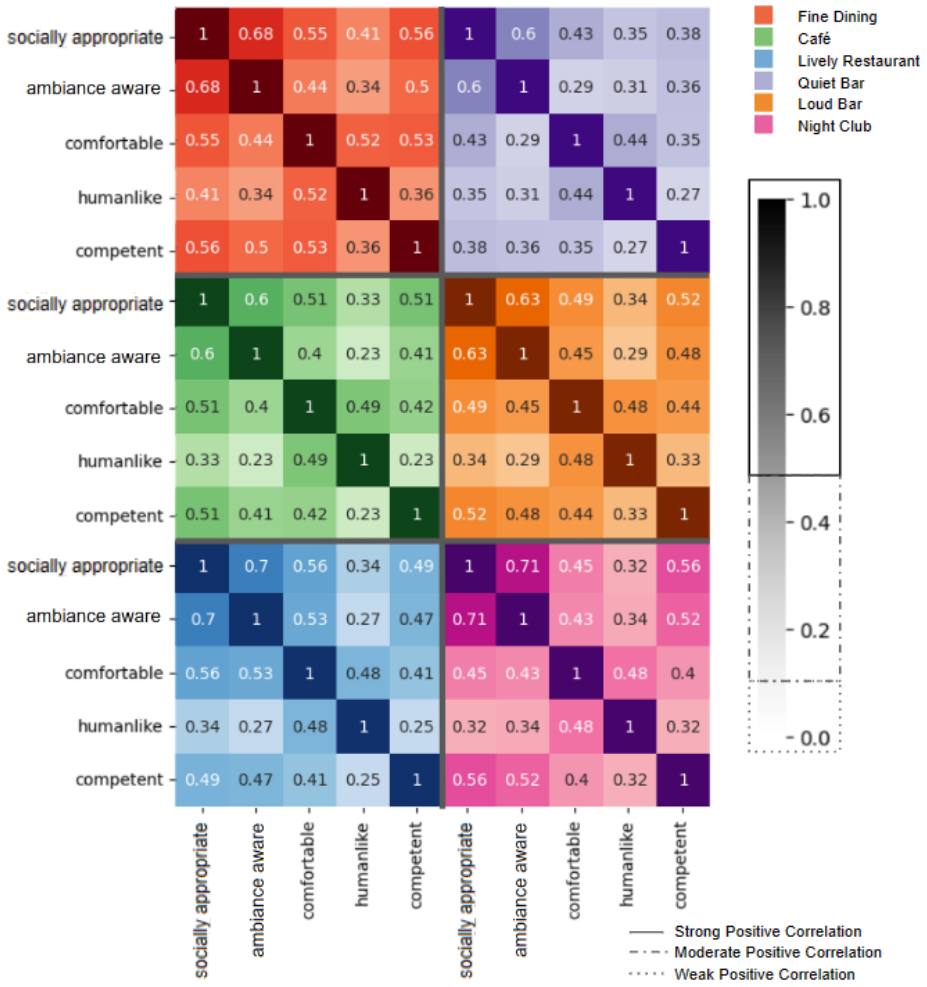

Figure 4.11: Pearson's R correlations between perception study responses for each ambiance.

Table 4.4: Chi-Square preference results

Ambiance	Count			χ^{22}	P-value	CI[3]
	Neutral	Joyful	'Lomb.'			
FIRST CHOICE						
Fine Dining	60	33	**27**	15.45	0.001**	13-32%
Lively Res.	**25**	55	38	11.51	0.003**	12-31%
Quiet Bar	52	49	**19**	15.49	0.001**	5-22%
Nightclub	**26**	58	36	15.49	0.001**	12-31%
LAST CHOICE						
Lively Res.	58	**25**	35	14.65	0.001**	12-31%

1. **Bold** values highlight the voices selected at a significantly different ratio.
2. DF = 2, N = 120 (118 for Lively Restaurant)
3. Confidence interval on selection proportion is reported only for the selected voice at a significantly different ratio.
4. Only significant values are reported.

positive correlations are seen between social appropriateness, competency, and comfort for fine dining and the café, with competency in the nightclub and loud bar and with comfort in the lively restaurant. Comfort strongly correlates with competency and human-likeness for fine dining and ambiance awareness in the lively restaurant. Lastly, ambiance awareness shows strong positive correlations with competency in the fine dining and the nightclub, as well as comfort in the lively restaurant. All other correlations are moderately positive, which overall tells us that by increasing social and ambiance appropriateness, we can improve perceptions of social and overall intelligence in all factors.

4.6 Discussion

We aimed to better understand how to increase perceived robot intelligence through selection of ambient adapted voices, specifically using non-loudness features, in opposition to previous literature.

We first developed a protocol for collecting ambiance-modified voices through Zoom to mitigate the difficulties with in-the-wild collection of vocal features in noisy ambiances. We found that we were able to replicate our desired ambiances to a reasonable degree through changes in lighting, sound, and images, as confirmed through both our validation study and the ambiance ratings in our perception study. These results are especially exciting as they suggest that simple, low-cost virtual simulations of ambient environments could be a suitable means for data collection toward usage on real robots. Further studies should be conducted to compare the results of simulated versus in-the-wild contexts.

We proposed clustering the voice features to extract primary voice styles, as opposed to a brute-force method of creating a separate voice for each ambiance. This method requires less data and computational intensity while being more transparent and interpretable. Testing robot voices selected based off these clusters, we found that the presence of a voice style in

Table 4.5: Results for difference of mean Statistical tests

Ambiance	Neutral		Joyful		'Lombard'		DF	N	F	P-value	
	μ	σ	μ	σ	μ	σ				One-Way Anova	Tukey HSD
SOCIALLY APPROPRIATE											
Fine Dining	**5.30**	**1.49**	4.30	1.73	3.97	1.66	2	120	22.58	< 0.001***	< 0.001***
Lively Restaurant	4.48	1.63	**5.68**	**1.14**	4.82	1.59	2	118	18.93	< 0.001***	< 0.001***
Noisy Bar	4.41	1.54	**5.29**	**1.17**	4.61	1.47	2	120	20.07	< 0.001***	< 0.001***
Nightclub	3.88	1.54	**5.46**	**1.26**	3.28	1.87	2	120	38.28	< 0.001***	< 0.001***
AMBIANCE AWARE											
Fine Dining	**5.01**	**1.54**	4.26	1.65	4.22	1.63	2	120	9.43	< 0.001***	< 0.001***
Lively Restaurant	4.31	1.42	**5.29**	**1.26**	4.58	1.58	2	118	13.92	< 0.001***	< 0.001***
Noisy Bar	4.41	1.42	**5.29**	**1.17**	4.64	1.48	2	120	12.46	< 0.001***	0.001**
Nightclub	3.73	1.59	**5.22**	**1.27**	3.96	1.54	2	120	35.25	< 0.001***	< 0.001***
COMFORT											
Nightclub	4.35	1.45	**5.09**	**1.40**	4.41	1.56	2	120	9.43	< 0.001***	< 0.001***
HUMAN-LIKENESS											
Nightclub	3.66	1.57	**4.37**	**1.59**	3.65	1.60	2	120	8.05	< 0.001***	0.002**
COMPETENCY											
Noisy Bar	5.32	1.43	**5.87**	**1.35**	5.36	1.45	2	120	7.91	< 0.001***	0.002**
Nightclub	5.13	1.37	**5.80**	**1.06**	5.08	1.43	2	120	11.46	< 0.001***	< 0.001***

1. **Bold** values highlight the preferred voice.
2. Due to space limitations, we only report those voices that were found to have a significantly different Likert rating than both other voices. In all cases, the post-hoc P-value was equal for all pairings with the preferred voice.

an ambiance from the Zoom study (Fig. 4.6) tended to match the preference of the matching voice style in that same ambiance in the on-robot study (Fig. 4.10). This provides evidence supporting our overall procedure.

Our results indicate that the Neutral voice (default on Pepper) was significantly NOT preferred in loud, high-energy environments (Table 4.4). Instead, a higher-pitched voice with high pitch range and speed was preferred. Surprisingly, given both the literature on human vocal adaptation and our human voice analysis results, we did not see our curated "Lombard" voice outperforming Pepper's base Joyful voice in these loud environments. However, the designed "Lombard" voice tended to have higher comfort or social appropriateness ratings in the loud environments (loud bar, lively restaurant, nightclub) compared to the Neutral voice. One participant commented that the "Lombard" voice was "Robotic sounding but louder" in the nightclub setting. Since the robot volumes remained constant, this could suggest that the "Lombard" voice with increased pitch could indeed increase perceived loudness. Finally, it should be noted that our in-person study did *not* include any negative sentiment sentences. It remains to be tested whether the "Lombard" voice could outperform the Joyful Pepper voice in these contexts (e.g., "Sorry, we don't have that today."). As expected, the Neutral voice was found to be more socially appropriate and ambiance-aware in the quiet fine dining ambiance.

The lack of preference for the "Lombard" voice may have several explanations: first, the Lombard voice is produced by humans in loud environments to increase intelligibility. However, further research needs to be done to see if equally loud and intelligible voices are

actually preferred by the listener or are just an inevitable production of vocal force from the speaker. Participants commented that with the Lombard voice in the nightclub environment, "the interaction was like with a real server", but "could be more casual". Another possible explanation is the lack of flexibility in text-to-speech models. Perhaps only being able to modify speech rate and pitch with a slight selection of pitch range is not enough to accurately re-create Lombard speech. Spectral features, specifically spectral slope, which was found to be significantly different in loud environments, may play a key role in the acceptability of pressed Lombard voice. Finally, by inspecting Fig. 4.6, the preference for the joyful voice over Lombard voice may, in fact, be expected. In our Zoom study, the joyful voice style was used more often in all ambiances compared to the Lombard voice style. However, utterance-specific appropriateness may also be considered. For our perception study, all utterances were positive or neutral. It seems likely that a Lombard voice may outperform a joyful voice in loud ambiances if the utterance had a negative sentiment (e.g., "Sorry, we don't have that today.") and should be considered future work.

Overall, other than the café and quiet bar, including those ambiances where there was no significance in the selection of preferred server, the preferred voice saw a significant increase in social appropriateness and ambiance awareness. This suggests that carefully selecting the correct voice for an ambiance can improve human perception of socially intelligent robots. Additionally, Fig. 4.11 depicts strong correlations between an increase in socially appropriate voice and competency.

It is important to note the limitations of the aforementioned study. First, although we collected clear audio data for a variety of different virtual ambient environments, we could not control for how one would adapt their voice virtually versus in a real ambient environment. Second, we could not control for group dynamics in the user study. More specifically, how one would interact, perceive, or interpret Pepper in isolation may differ from when not in isolation. Finally, we were limited to Pepper's vocal features, and thus, our work was tailored to that of Pepper. Introducing a variety of social robots in the future would allow us to observe if embodiment impacts the results.

Chapter 5

A Text-to-Speech Model for Second-Language Listeners

We finally arrived at the question of how to adapt a TTS to improve comprehension for second-language speakers. The perception and production of certain L2 vowels, even those that exist in the speaker's first language, remains challenging for L2 speakers even at a high level of proficiency [159]. In English, differentiating between tense (long) and lax (short) vowels (e.g., beat vs. bit) can be particularly difficult for second-language speakers [70, 71, 160].

Often, first-language speakers consider that slowing down their speech rate or putting emphasis on a difficult sound by raising their pitch will help improve L2 comprehension [68]; however if vowel length or pitch is a primary marker for vowel differentiation, is this truly the ideal method?

As already mentioned in Chapter 2, while in the past vowel perception had been viewed as a local process [116], it has now become well-understood that the pitch and duration context surrounding vowels plays an important role in their perception [73]. The existence of such context effects opens the opportunity to use pitch and duration manipulations within and around the target sound strategically to bias the perception of difficult L2 sounds into their correct interpretation. Unfortunately, we lack a precise specification of how one should modify these cues to facilitate vowel comprehension.

To understand the contribution of pitch and duration of both a target word containing a tense or lax vowel and its proceeding context, in [78], we applied the data-driven psychophysical methodology of *reverse correlation*. We follow this up with a validation of these findings, making manual manipulations to a TTS. Finally, we built a TTS that automatically incorporates modifications to tense-lax vowels specifically designed to aid L2 speakers in their comprehension of these difficult vowels.

5.1 Reverse correlating context effects on L1 and L2 vowel perception

To explore the effect of pitch and duration on vowel perception, we used a phase-vocoder technique [161] to systematically vary the pitch and speech rate in phrases surrounding pairs of vowels/words that are difficult for L2 speakers: English (/i/-/ɪ/) and French (/u/-/y/), letting participants listen and judge which word they heard, and then used reverse correlation to reconstruct the prosodic profiles that biased the perception of these word pairs in one direction or the other. We did this both for isolated words (Word task) and for words embedded in sentences (Phrase task).

We first collected data from English-L1-French-L2 and French-L1-English-L2 speakers on the same French and English stimuli. This initial pairing was chosen as duration plays a vital role in English vowels, yet plays little role in distinguishing vowels in Parisian French [162–164]. Additionally, while pitched stress plays a vital role in English [165], stress in the French language is still not clearly defined and appears to be more rhythmic in nature [166].

We then extended the study to Mandarin-L1-English-L2 and Japanese-L1-English-L2 speakers with the same English stimuli as the previous groups. As French has both lesser influences of pitch and duration on the comprehension of vowels than in English, Mandarin, and Japanese were chosen to explore these aspects in languages that rely more extensively on these features for perception. Mandarin lacks vowel length contrasts, and it has been found that Mandarin speakers display reduced precision in determining vowel length in other languages [167]. Instead, Mandarin vowels are firmly defined by four discrete tones completely defined by changes in pitch over the vowel, where duration remains only a secondary cue (e.g., 汤tāng (soup) and 糖táng (sugar)). The addition of Japanese extends to speakers whose first language relies heavily on duration to differentiate vowels. All 5 vowels in Japanese contain a minimal pair of short/long vowels, which, when varied, can change the meaning of a word [168](e.g. かど/kado/ (corner) カード/ka:do/ (card). Moreover, Japanese has a pitch accent distinct from, but not dissimilar to, English [169].

5.1.1 Stimulus generation

Vowel selection and word

In English, we selected the pair of tense/lax vowels /i/ and /ɪ/ that are known to be difficult for English L2 speakers both in perception and in production [71]. From the point of view of French-L1 speakers, the vowel /i/, such as in “beat” /bit/, is a high, tense vowel and exists in both French and English. In contrast, the vowel /ɪ/, such as in “bit” /bɪt/, is lower and lax and does not exist within the French lexicon. Often, French-L1 speakers learning English will replace /ɪ/ by /i/, for example saying “sheep” [ʃip] when they mean to say “ship” /ʃɪp/ [71]. Similarly, [70] found that Japanese speakers will form the same vowel substitution, i.e., substitute the tense /i/ for the lax /ɪ/. In Japanese, as in English

and French, /i/ is a high front vowel, and, as in French, /ɪ/ is non-existent. Mandarin-L1 speakers of English face a similar problem. While Mandarin contains the vowel /ɪ/, Mandarin speakers often lack the ability to distinguish between /ɪ/ and /i/ as, in Mandarin, the length difference between these two vowels does not alter their meaning. In many cases, Mandarin speakers will replace both vowels in "sheep" [ʃip] and "ship" /ʃɪp/ with the Mandarin vowel /ɪ/, which requires a higher and more frontal position of the tongue [160]. We selected target words that differ only by these vowels to maintain consistency in confounding factors from consonant proximity. The selected words were "pill" /pɪl/ and "peel" /pil/.

For French, we selected the vowels /u/ and /y/. Once again, this pair of vowels is known to be difficult for English-L1-French-L2 speakers in both perception and production [170, 171]. The vowel /u/, such as in "fou" /fu/ (mad), is a high, back vowel and exists in both French and English. In contrast, the vowel /y/, such as in "fût" /fy/ (cask), is a high, front vowel and does not exist within the English lexicon. English speakers tend to overcompensate for the lack of /y/ vowel in their language production and often replace /u/ with /y/, such as mispronouncing "beaucoup" /boku/ (a lot) with the rather immodest phrase [boky] [172]. Once again, we selected words that differ only by these vowels to maintain consistency in confounding factors from consonant proximity. The selected words were "pull" /pyl/ and "poule" /pul/.

Phrase stimuli

To embed target words in a larger context, we used the phrase "I heard them say" (FR: "je l'ai entendu dire") preceding the word. This was selected in an attempt to avoid that a semantic context could bias the interpretation of the target word in the direction, e.g., of the most frequent alternative. The phrases were generated using the Hugging Face interface for CoquiXTTS[1]. The language was set to English for the English phrase and French for the French phrase. An L1 male reference voice was provided, and no other modifications were made to the TTS settings.

Stimulus manipulation: vowel ambiguity

To avoid response bias, reverse correlation experiments most often employ a 2-interval one alternative methodology, for example, asking the participant: "which of the two stimuli is more like *a target*". However, for sound-based experiments, the experiment length nearly doubles with the addition of a second stimulus. Comparatively, a 1-interval task ("which of these 2 words did you hear") allows for nearly twice as many trials in the same amount of time. Yet, because reverse correlation presents the participant with random manipulations of an original word/phrase in the hope that noise will sway their decision one way or another,

[1] https://huggingface.co/spaces/coqui/xtts

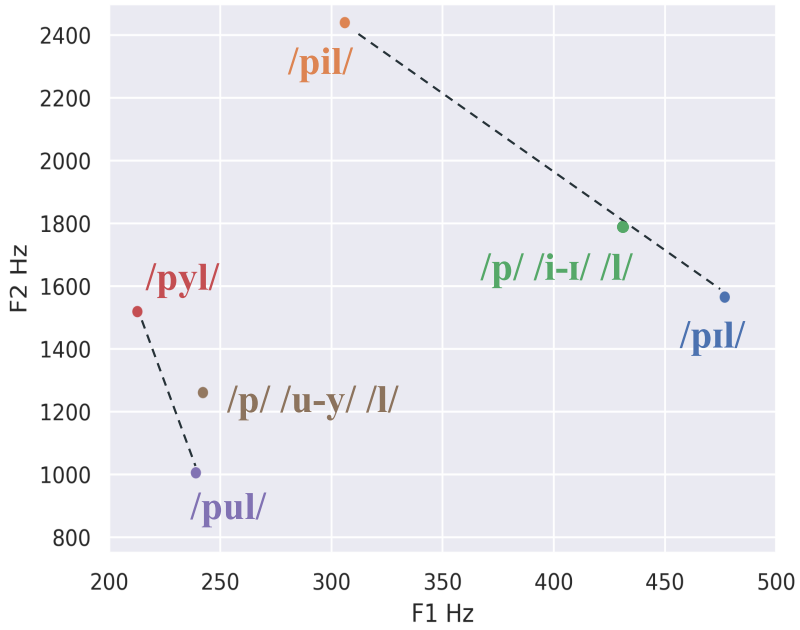

Figure 5.1: F1-F2 plot of initial and final formants for ambiguous vowels /u-y/ and /i-I/.

the 1-interval design presents the risk that the modifications made to the phrase will not be enough to sway the participants between responses; if not, there will be a strong bias towards the original word, limiting the number of trials that can be analyzed. Another problem with 1-interval tasks is that we would need two sets of phrases, one for each word.

To mitigate this bias in our 1-interval, 2-alternative task method, we generated morphed sounds that were perceptually intermediate between each of the vowel pair: /i/-/ɪ/ and /u/-/y/. To do this, we extracted the vowels from each of the 4 target words, making sure to split and merge on zero-crossing points, and estimated the frequencies of the first two formants, F1 and F2, for each vowel. We then used the Whisper automatic speech recognition (ASR) [145][2] to estimate the most probable interpretation at each of a series of steps interpolated between the formants of each alternative and selected the point that Whisper judged most ambiguous.

In more detail, we modified Whisper to extract the log-probabilities that a processed audio clip contains each of the target words for a given language rather than simply giving the most likely prediction. A two dimensional search grid was generated using the origin F1 and F2 as follows: /i/ F1: 305.89Hz, F2: 2440.77Hz to /ɪ/ F1: 476.85Hz, F2: 1565.45Hz, stepsize = 10Hz; /u/ F1: 238.90Hz, F2: 1005.67Hz to /y/ F1: 212.61Hz, F2: 1519.49Hz. At each iteration, Praat [173] was used to extract and modify the formants, then re-synthesize the word to be processed by the ASR model. This was completed starting from both vowels, i.e. in English moving both from /pil/ (peel) to /pɪl/ (pill) and from /pɪl/ (pill) to /pil/

[2] v20231117, medium multilingual

(peel), searching for the formant pairing with the smallest difference in log-probabilities for the two target words (i.e. most ambiguous). The resulting formants can be seen in Fig. 5.1 with the final /i/-/ɪ/ vowel having F1: 436.77Hz, F2: 1722.68Hz (synthesized from "peel" /pil/), and the final /u/-/y/ vowel having F1: 238.13Hz, F2: 1258.68Hz (synthesized from "poule" /pul/). These ambiguous words served as base stimuli in the Word task (see Experimental Procedure, below) and were then inserted manually in their original phrase (at a zero-crossing 120ms after the end of the last word "say/dire") to serve as base stimuli for the Phrase task.

Stimulus manipulation: randomization

Next, we generated the reverse-correlation stimuli from these ambiguous base sounds using the open-source CLEESE toolbox [174], a voice-transformation toolbox that creates random fluctuations around an audio file's original contour of pitch and speech rate. The pitch contour of the recordings was artificially flattened to a constant 120Hz. We then transformed the stimuli by randomly manipulating their pitch and duration independently, beginning with pitch, then duration, in n successive windows of 100ms (n=4 for words; n=13 for phrases), each of which with a factor sampled from a normal distribution (pitch: μ=0, σ=100 cents (i.e. 1 semitone); duration: μ=0%, σ=100% (i.e. doubling or halving the window's duration); both distributions clipped at $\pm 2\sigma$). These values were chosen so as to cover the range observed in naturally produced utterances and were linearly interpolated between successive time points to ensure a natural-sounding transformation.

5.1.2 Experimental procedure

Reverse correlation is an experimental paradigm aimed at discovering the signal features that govern a participant's judgment by analyzing their responses to large sets of stimuli whose acoustic characteristics have been systematically manipulated [175]. Specifically, we presented participants with a series of 250 trials, each consisting of a single base recording containing one of our ambiguous target words. For each trial, the base recording was manipulated as previously described with a different random profile of pitch and speech rate, and participants were asked which of two target words they heard (1-interval, 2-alternative forced choice). Responses were then analyzed to reconstruct the prosodic profile that maximizes the likelihood of responding to one option or the other (for a review, see [128]).

This experiment included 4 different conditions, randomized between participants: (A) two involving random manipulations of the isolated target word (English or French), aimed at establishing a baseline for the intrinsic pitch and rate of the two alternatives, and (B) two involving manipulations of phrases containing the target word, aimed at uncovering extrinsic acoustic context effects. Participants recruited online had the option to participate in more than one condition, in which case the order was randomized across language and type of stimuli (word and phrase).

In each condition and before the reverse correlation task, we collected information about the participant's native language, age, gender, and self-rated English proficiency. Participants then successively listened to all trials (each played only once) and responded which of the target words they heard. The order of response options (e.g., "pill/peel" or "peel/pill") was randomized across trials and participants. Each condition lasted, on average, 15 minutes.

Participants

N=160 participants took part in the study: N=54 French-L1 speakers (female: 24, M=32.4yo ± 9.7), N=60 English-L1 speakers (female: 33, M=30.0yo ± 10.7), N=30 Mandarin-L1 speakers (female: 15, M=22.2yo ± 5.6), and N=16 Japanese-L1 speakers (female: 5, M=29.9yo ± 12.1). The participants were recruited via the online Prolific platform.

L2 participants had a self-rated English proficiency ranging from 2-5 (1: no proficiency, 5: fluent) with a mode of 4. English participants were recruited primarily from anglophone Canada, French participants from France, Mandarin within China, and Japanese from Japan. As such, the L2 speakers are immersed in their native language rather than their L2.

English and French participants were recruited for all English/French Word/Phrase tasks (4 conditions). The Japanese and Mandarin participants were only recruited for the English vowel tasks (2 conditions: Word and Phrase). Each condition (1-word tasks, 1 phrase tasks) was carried out with n=25 speakers for each L1 aside from Japanese. For Japanese participants, we had n=14 for the word task and n=13 for the phrase task due to difficulties finding Japanese participants. Although conditions were randomized across participants, participants were also left the option to participate in more than one condition - in the following, we treat all samples as independent regardless of repeated measures.

All participants provided their informed consent and were compensated financially for their time at a standard rate. The procedure was approved by the SFU-REB.

5.1.3 Results

Validation of ambiguity

We first explored how successful we were at creating an ambiguous vowel to confirm the validity of the PRAAT/Whisper approach. Because random manipulations were centered on zero, we expected a 50% response rate for the alternative options. For the *English Word* condition, English-L1 speakers answered "peel" 52% of the time, French-L1 speakers 54% of the time, Mandarin-L1 speakers 53% of the time, and Japanese 52%. For the *English Phrase* condition, English-L1 speakers responded with 59% "peel", the French-L1 speakers with 54%, Mandarin-L1 speakers with 54% and Japanese-L1 56%. For the *French Word* condition, English-L1 speakers responded "poule" 64% of the time and French-L1 speakers

70% of the time. For the *French Phrase* condition, English-L1 speakers responded with 59% "poule", while French-L1 speakers' responses were 66.6% "poule".

In sum, all stimuli appeared sufficiently ambiguous, although to a lesser degree for native speakers and for French. Informal discussions with participants concurred with these results.

French reverse correlation

The first subset of results concerns how L1 (French) and L2 (English) speakers perceived French-language stimuli, either as single words (*"pull"* vs. *"poule"*) or complete sentences (ex. *"je l'ai entendu dire poule").*

Analysis procedure: We computed first-order kernels from reverse-correlation data using the *classification image* method [128] for each participant. This is done by computing the average random pitch and speech rate transformation profile of the recordings classified as one response option (e.g., "pill") and subtracting it from the average profile of the recordings classified as the other option (e.g., "peel"). The kernels are then normalized by dividing them by the root-mean-square sum of their values. This resulted in two dim=4 (word) and dim=13 (phrase) vectors of pitch and rate of speech values for each participant. These represent the pitch and speech rate transformations that should be applied to a given base word to increase the likelihood of recognizing one target word or the other. To analyze the differences between the two kernels within participants, we compute paired t-tests at every time point, i.e., every 100ms.

Words: Our prediction was that French-L1 (FL1) speakers' perception of /y/, a high/front vowel, would be driven by a higher pitch and faster speech rate compared to /u/. Our results confirm this (Fig.5.2) for speech rate (FL1 - 0.1s: $t(25)=2.80$, $p=.010$; 0.2s: $t(25)=3.52$, $p=.002$), and pitch (at 0s and 0.1s, albeit non-statistically). French-L2 (EL1) shares the same pattern of data, with a non-statistical pitch increase at t=0.1s and faster speech rate at 0.1s: $t(25)=4.57$, $p<.001$; and 0.2s: $t(25)=2.73$, $p=.012$. Although statistically weak, this pattern of results (higher pitch/faster rate for /y/) is confirmed by segments located on the target word in the phrase kernels (see below).

Phrases: Phrase reverse-correlation data first confirms, this time significantly, that, within the target word, hearing /y/ is driven by higher pitch in FL1 (1s: $t(25)=-2.63$, $p=.015$) and faster speech rate in both FL1 and EL1 speakers (FL1 - 1.0s: $t(25)=2.99$, $p=.006$; EL1 - 1.0s: $t(25)=4.04$, $p<.001$, 1.1s: $t(25)=3.71$, $p=.001$). Second, regarding the surrounding context of the word, the literature predicts the existence of contextual contrast effects, i.e., that /y/ is driven by *lower* pitch and speech rate before the target word. In actuality, our data reveals a mix of contrastive and congruent contextual influences: pitch (Fig. 5.3-left) is not associated with any contrastive effect, but rather a proximal congruent effect 200-300ms pre-target (FL1- 0.7s: $t(25)=-2.18$, $p=.040$; EL1- 0.8s: $t(25)=-4.05$, $p<.001$), i.e. an increase of pitch immediately before the target word biases the response towards the higher-pitch alternative /y/. Speech rate (Fig. 5.3-right) exhibits the expected long-term (distal)

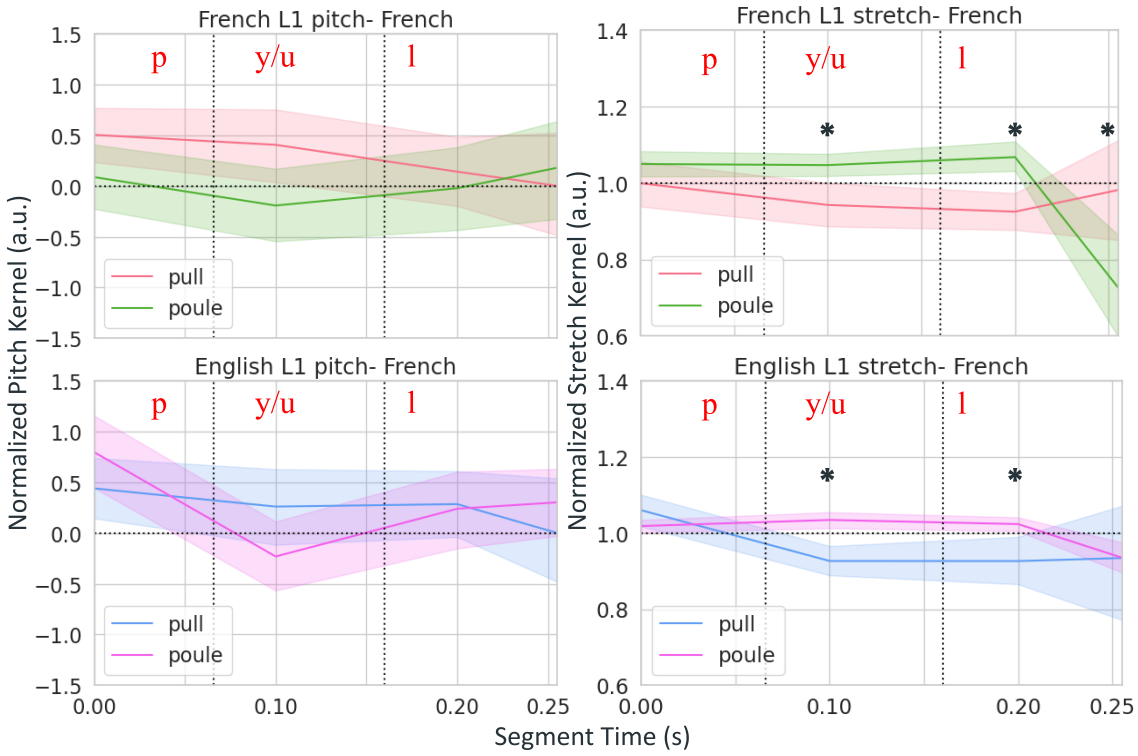

Figure 5.2: **French words reverse-correlation results:** Pitch (left) and speech rate (right) kernels for French-L1 (top) and L2 (bottom) speakers for the French words "pull" and "poule", presented in isolation. In all figures, colored areas mark 95% confidence intervals on the mean, and * marks time segments that differ statistically at $\alpha = 0.05$. Smaller values for the stretch kernel mean shorter duration, i.e. *faster* speech rate, and for the pitch kernel lower pitch.

contrastive effect (FL1 - 0.2s: t(25)=-3.72, p=.001, 0.3s: t(25)=-3.86, p=.001, 0.6s: t(25)=-2.89, p=.008, 0.7s: t(25)=-2.24, p=.018; EL1 - 0.1s: t(25)=-3.35, p=.003, 0.2s: t(25)=-2.53, p=.018, 0.3s: t(25)=-4.12, p<.001, 0.4s: t(25)=-2.99, p=.006, 0.6s: t(25)=-3.46, p=.002), where a slower speech rate at the beginning of a phrase and a proximal congruent effect 100ms pre-target (FL1 - 0.9s: t(25)=3.25, p=.003; EL1 - 0.9s: t(25)=3.76, p=.001) biases the response towards the faster alternative /y/, resulting in an overall scissor-shape profile. This pattern of result was remarkably conserved in L2 speakers (Fig. 5.3-bottom).

English reverse correlation

A second subset of results concerns how L1 (English) and L2 (French, Chinese, Japanese) speakers perceived English-language stimuli, either as single words (*"peel"* vs *"pill"*) or complete sentences (ex. *"I heard them say pill"*). ccepting either extended abstracts (2-pag **Words:** We predicted that EL1 speakers' perception of /i/, a high/tense vowel, would be driven by a higher pitch and slower speech rate compared to /ɪ/. Yet, in EL1, our results do not confirm this prediction (Fig.5.4, top) as no significant differences are found between the kernels. English L2 speakers do, however, show the expected effect on speech rate -

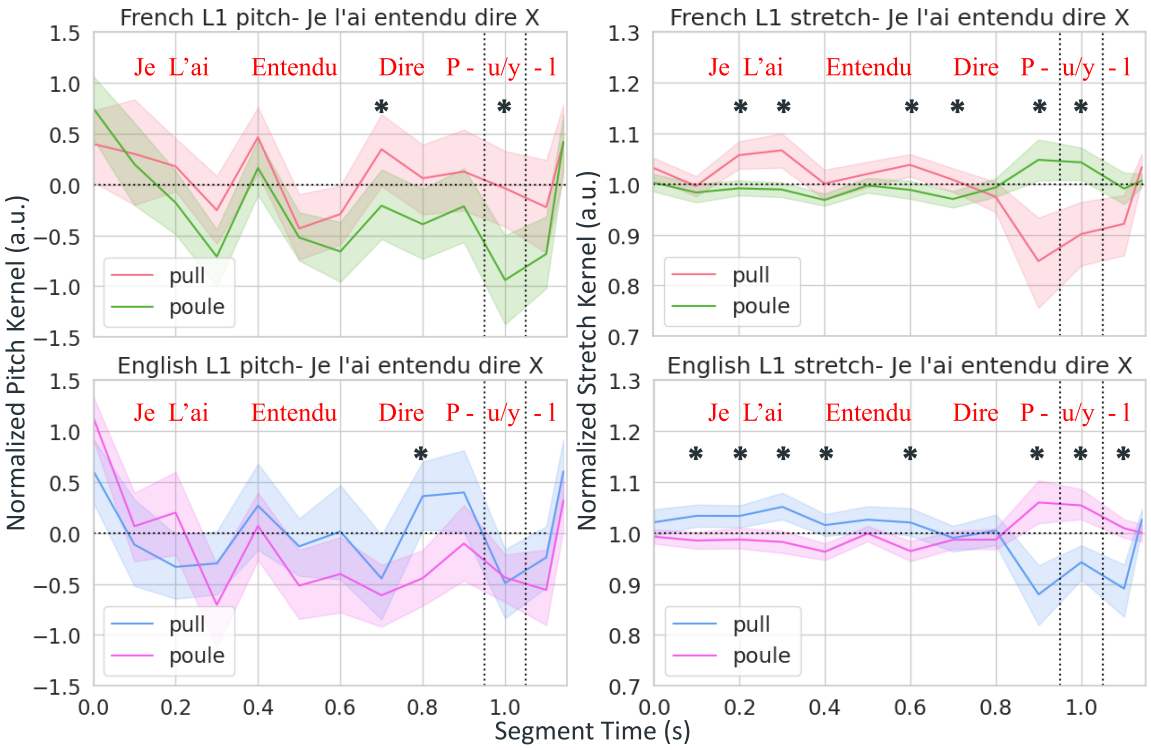

Figure 5.3: **French phrase reverse-correlation results:** Pitch (left) and speech rate (right) kernels for French-L1 (top) and L2 (bottom) speakers for the French phrases containing "pull" and "poule". Smaller values for the stretch kernel mean shorter duration, i.e. *faster* speech rate, and for the pitch kernel lower pitch.

although not on pitch. French L1 speakers display a significantly faster rate of speech for /ɪ/ (0.2s: t(25)=3.13, p=.005) (Fig.5.4, bottom). Mandarin speakers (ML1) show same expected lengthening in /i/ and shortening in /ɪ/ (0.1s: t(25)=3.58, p=.001) (Fig.5.5- top right) as in FL1, but earlier in the vowel than for FL1 speakers. Lastly, for Japanese speakers (JL1), we see a strong shortening across the entire vowel for /ɪ/ (0.1s: t(14)=5.01, p=<.001; 0.2s: t(14)=6.95, p<.001) (Fig.5.5- bottom right). None of the second-language speakers show any significant effect for pitch,

Phrases: Regarding the surrounding context of the word, the literature predicts the existence of contextual contrast effects, i.e., that /ɪ/ is driven by *lower* pitch and *faster* speech rate before the target word. In actuality, like in the French phrase reverse correlation, our data reveals a mix of contrastive and congruent contextual influences.

For pitch, contrary to theoretical predictions, /ɪ/ is driven by a significantly higher pitch (0.9s: t(25)=-2.69, p=.013) for EL1 speakers. Moreover, pitch shows contrastive effects within the phrase, both distally (0.1s: t(25)=3.23, p=.003, 0.3s: t(25)=3.68, p=.001) and in the immediate proximity of the target word (0.8s: t(25)=2.91, p=.007) (Fig. 5.6- top left). For English L2 speakers, as in the word, we do not see significant differences in the target word for pitch, aside from Japanese speakers that prefer a slightly lower pitch in

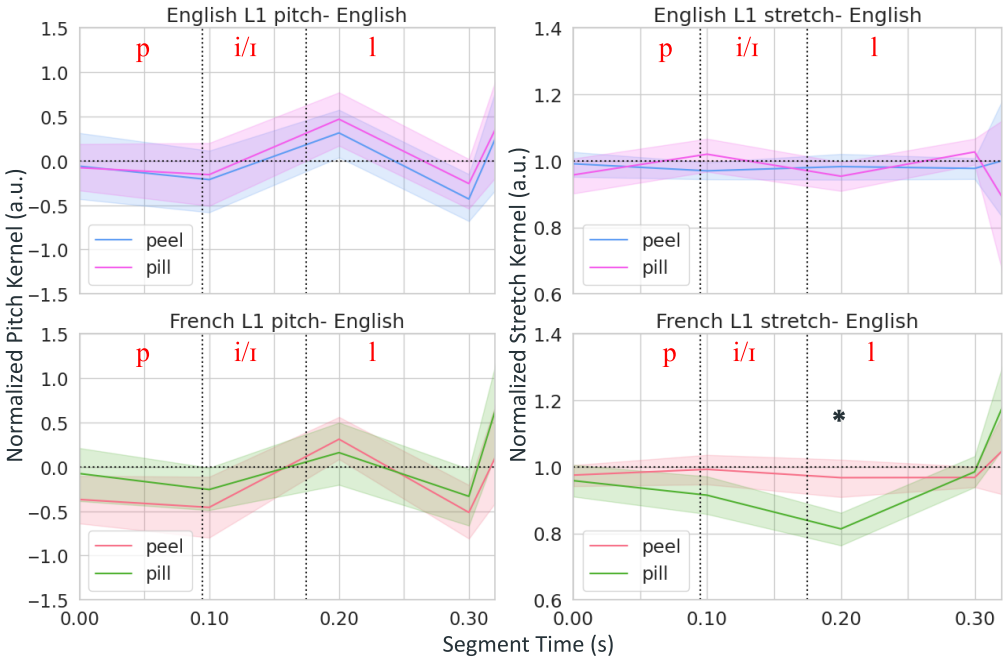

Figure 5.4: **English words reverse-correlation results (EL1, FL1):** Pitch (left) and speech rate (right) kernels for English-L1 (top) and L2 (bottom) speakers for the English words "peel" and "pill". Smaller values for the stretch kernel mean shorter duration, i.e. *faster* speech rate, and for the pitch kernel lower pitch.

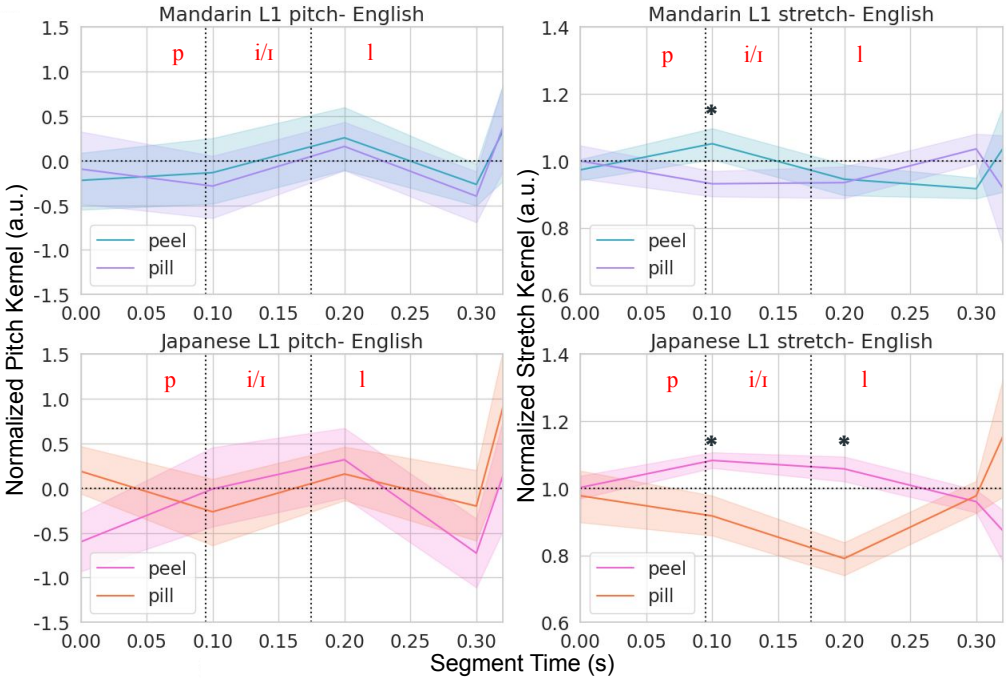

Figure 5.5: **English words reverse-correlation results, continued (ML1, JL1):** Pitch (left) and speech rate (right) kernels for Mandarin-L1 (top) and Japanese-L1 (bottom) speakers for the English words "peel" and "pill". Smaller values for the stretch kernel mean shorter duration, i.e. *faster* speech rate, and for the pitch kernel lower pitch.

"peel" following the vowel (1.0s: t(13)=-2.66, p=.019) (Fig. 5.7- bottom left). Instead, there appears to be a preferred pitch pattern over the entire phrase that is interestingly similar across FL1 and ML1 speakers, especially within the target word.

EL1 speech rate (Fig. 5.6-top right) exhibits the expected long-term (distal) contrastive effect (0.1s: t(25)=-2.52, p=.018, 0.2s: t(25)=-2.31, p=.030, 0.4s: t(25)=-3.47, p=.002, 0.5s: t(25)=-2.34, p=.028), where a slower speech rate at the beginning of a phrase and a proximal congruent effect up to 200ms pre-target (EL1: 0.7s: t(25)=3.77, p=.001, 0.8s: t(25)=5.98, p<.001), biases the response towards the faster alternative /ɪ/, resulting in an overall scissor-shape profile. However, the rate difference within the target vowel is not significant for EL1. This is not due to the speech rate at the proximal segment, which moves in the expected direction, but rather to the unexpected reversal from the expected speech rate within the vowel.

Remarkably, we see the scissor shaped pattern of speech-rate results was conserved in all groups of L2 speakers. As in the word task, we confirm the expected difference in speech rate, i.e. a shorter /ɪ/ (0.9s: t(25)=2.15, p=.042, 1.0s: t(25)=3.61, p=.001) for FL1 speakers. The FL1 show the same scissor-shape pattern we saw for EL1 speakers (Fig. 5.6-bottom right), with distal contrastive effects (0.1s: t(25)=-3.02,p=.006, 0.3s: t(25)=-3.53,p=.002, 0.4s: t(25)=-3.07, p=.005, 0.5s: t(25)=-2.75, p=.011), and a strong proximal congruent effect (0.8s: t(25)=4.25, p<.001). ML1 speakers show the same scissor pattern with distal contrastive effects and proximal congruent effects (Fig. 5.7- top right) that we see in both the EL1 and FL1 speakers (0.1s: t(25)=-3.39, p=.002, 0.2s: t(25)=-2.11, p=.044, 0.3s: t(25)=-3.01, p=.006, 0.4s: t(25)=-3.65, p=.001, 0.5s: t(25)=-5.66, p<.001, 0.6s: t(25)=-2.28, p=.031, 0.8s: t(25)=6.36, p<.001, 0.9s: t(25)=5.43, p<.001, 1.0s: t(25)=2.44, p=.022). Once again, we see the same pattern in the JL1 group as we do in all other groups with significance across the entire phrase as in ML1 (Fig. 5.7- bottom right), despite the smaller sample size (0.1s: t(13)=-5.25, p<.001, 0.2s: t(13)=-2.82, p=.014, 0.3s: t(13)=-2.63, p=.021, 0.4s: t(13)=-3.02, p=.010, 0.5s: t(13)=-4.51, p=.001, 0.6s: t(13)=-2.14, p=.052, 0.8s: t(13)=4.01, p=.001, 0.9s: t(13)=4.08, p=.001, 1.0s: t(13)=3.39, p=.005).

5.1.4 Discussion

This experiment is, to the best of our knowledge, the first to apply a reverse correlation paradigm to examine how pitch and speech rate (i.e., traditionally prosodic cues) influence the perception of vowels in L1 and L2, both in isolated sounds and embedded in phrases (for a related recent study on how these cues influence the perception of word boundaries, see [176]).

Results in the French and English Word conditions (Figures 5.2, 5.4 and 5.5) show that reverse correlation can uncover intrinsic pitch and speech rate characteristics of vowels. Strikingly, we saw the same profile repeated through the experiments, adhering to the known characteristics of the vowel in all cases across four languages, except for the English

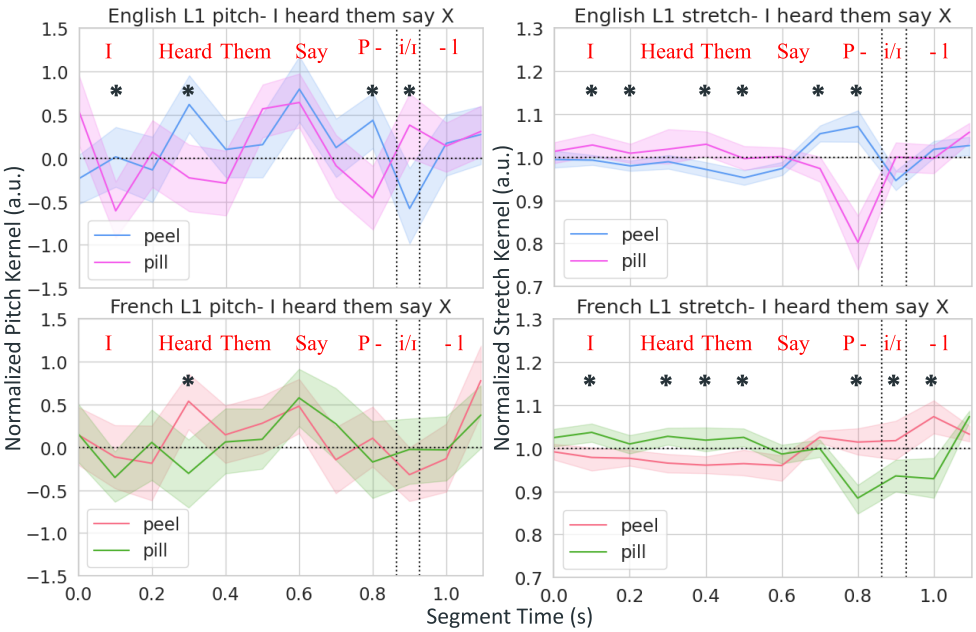

Figure 5.6: **English phrase reverse-correlation results (EL1, FL1)**: Pitch (left) and speech rate (right) kernels for English-L1 (top) and L2 (bottom) speakers for the English phrases containing "peel" and "pill". Smaller values for the stretch kernel mean shorter duration, i.e. *faster* speech rate, and for the pitch kernel lower pitch.

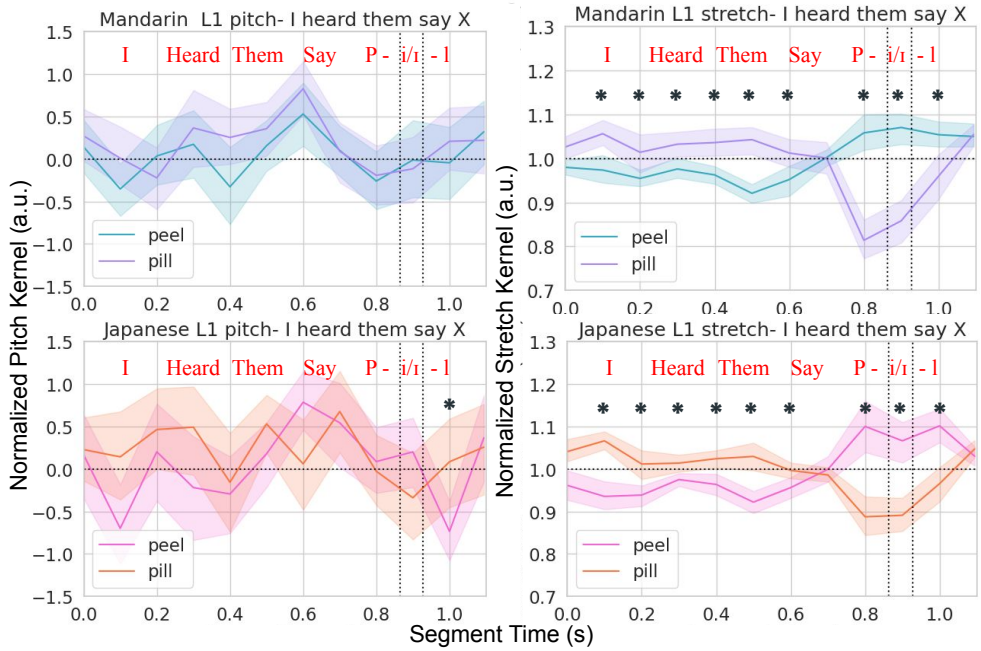

Figure 5.7: **English phrase reverse-correlation results, continued (ML1, JL1):** Pitch (left) and speech rate (right) kernels for Mandarin-L1 (top) and Japanese-L1 (bottom) speakers for the English phrases containing "peel" and "pill". Smaller values for the stretch kernel mean shorter duration, i.e. *faster* speech rate, and for the pitch kernel lower pitch.

vowel for L1 speakers. In this case, although /pɪl/ is generally considered to be longer and have a higher intrinsic pitch than /pil/, they are both high vowels, and the reversal of pitch characteristics is not particularly surprising. For speech rate, it may be that EL1 speakers put more weight on timbre cues than speech rate cues, whereas the L2 speakers still rely heavily on prosody to disambiguate the foreign timbre.

Results in the Phrase conditions (Figures 5.3, 5.6 and 5.7) provide a detailed account of how prosodic contextual effects may bias the perception of vowels when they're embedded in a phrase. We were able to replicate the predicted contrastive context effects in speech rate and weakly in pitch. Reverse correlation, however, helped us to uncover fine-tuned information on these effects. In particular, the effects are distally contrastive but become congruent proximal to the word, with a tipping point around 200ms before the target in French, Mandarin, and Japanese-L1 and 300ms before the target in English-L1. In general, we see weaker effects for pitch than speech rate, especially within the French phrase, and in general, the EL1 speakers saw more exaggerated pitch profiles in both languages.

Additionally, we found that L2 speakers could map remarkably well to the mental representation of L1 speakers both in terms of the intrinsic characteristics of the vowel and the contrastive effects of the context, in some cases even more significantly than L1 speakers. One reason for this may be that L1 speakers already are at ceiling performance when discriminating vowels based on spectral cues and do not typically rely on additional cues from the context, while context pitch and speech rate appear to constitute robust - and potentially determinant - cues for vowel perception in L2 speakers who may lack sufficiently resolute spectral representations.

Finally, it is also worth noting that phoneticians often concentrate on examining one single cue at a time to deeply understand its significance for perception. This, however, is not how humans perceive sound, and reverse correlation offers a methodology to explore combinations of cues at once, producing results that may be more akin to everyday human perception.

Overall, the most prominent effect found here is the scissor-shaped pattern of speech rate. That pattern was shown to bias vowel perception in both L1 and 3 samples of L2 speakers for both French and English vowel pairs. It was replicated with a strikingly similar shape in all experiments. In the rest of this chapter, we therefore focus on this perceptive mechanism, first to validate that this has a behavioural effect in a wider selection of vowels (Section 5.2, below) and, second, to use it as a strategy to improve L2 clarity in a state-of-art TTS architecture (Section 5.3).

5.2 Validation of Durational Vowel Control

Reverse correlation allows us to uncover psychoacoustic mechanisms that we use unconsciously to differentiate sounds. Our results suggested a mechanism to improve comprehen-

sion when the listener struggles to use the primary formant cues, such as with L2 speakers [159]. However, we need to confirm that these mechanisms translate to macroscopic behavioural changes (i.e., change perceptual responses) through validation experiments. First, we found a strong effect of duration within the target word and a weaker effect within the context. Still, we do not know the exact amount of duration change required to elicit the desired effects, i.e., a change in perception between a minimal tense/lax pair. In the following, we systematically manipulated the scissor manipulation's intensity to establish perception thresholds. Second, reverse correlation kernels have shown effects spanning the complete phrase and vowel but do not provide information on their respective strengths. In the following, we explore duration changes in the word and context, the context only, and the word only. Third, our previous results theoretically only hold for the two vowel pairs investigated. In the following, we chose to focus on English, as English has more clearly defined long and short (tense/lax) vowels that do not rely explicitly on surrounding consonants to determine their duration, and introduce another tense/lax vowel pair: "full" (/fʊl/) and "fool" (/ful/). Finally, while reverse correlation was conducted on manipulated ambiguous vowels for practical reasons (avoiding bias in a 1-interval task), we do not know how they extend to more ecological situations. In the following, the vowels are no longer ambiguous, allowing us to better understand and confirm whether L1 speakers rely more strongly on timbre cues while L2 rely on prosody.

5.2.1 Stimulus generation

We generated phrase stimuli (e.g. *"I heard them say fool"*) that incorporated the "scissor-shape" profile of speech-rate uncovered in the previous experiment. To do this, we again use Matcha-TTS [88] (see Chapter 3), which has phoneme level duration control and allows us to manipulate the speech rate at specific points of the phrase.

The fact that the reverse correlation kernels in the previous Phrase results (Fig. 5.6 and 5.7) had smaller weights (the y axis of the figures respresent the kernel weights) outside than within the target word suggests that the duration change required in the context to elicit a perception effect should be larger than in the word (kernel amplitudes are roughly equivalent to sensitivity [128]). However, we found that large duration changes in the context resulted in a highly unnatural speech rate that would not be applicable in real-world use cases. Therefore, we opted to make a smaller duration change to the context than within the target word. Within the word, we tested a duration change ranging from 0.5x speed to 2.0x speed at increments of 0.2, both increasing and decreasing from 1.0x speed, resulting in 11 different stretch manipulations. The corresponding context duration change ranged from 0.67x to 1.5x speed at increments of 0.1, both increasing and decreasing from 1.0x speed. Context and word duration were applied in opposite directions, e.g., when the word was 2.0x stretched, the context was compressed at 0.67x simultaneously. To explore the contribution of duration changes in the word and the context individually, we tested three

conditions: duration manipulations on the context and the word, the word only, and the context only. The duration changes for the word-only and context-only conditions were the same as when the word and context were combined.

Using Matcha-TTS, these modifications were made by hand by applying an array of the same length as the phonemized phrase. This method has been shown to result in natural and effective duration changes in TTS systems containing a phonemizer [177]. In this array, the phonemes up until the pause before the target word contained the context multiplier, and phonemes beginning at the space before the target word contained the word multiplier. This *clarity duration* multiplier is applied just after the base speech rate multiplier, where the base rate multiplier is an array the same length as the phonemized phrase containing only the speech rate provided at synthesis (in our case, 0.75). The Matcha-TTS text encoder returns x_μ: the average output of the encoder, $\log w$: the log duration predicted by the duration predictor, and x_{mask}: the mask for the text input. The clarity array *carray* is applied via Hadamard product to the array resulting from the Hadamard product of x_{mask}, the exponential of $\log w$ and the base speech rate multiplier *speechrate* (Algorithm 5.1). The resulting y_{lengths} and $y_{\text{max\_length}}$ are then used to calculate the attention alignment map.

$$\begin{aligned} w &= e^{\log w} \odot x_{\text{mask}} \\ w_{\text{ceil}} &= (\lceil w \rceil \odot \text{speechrate}) \odot \text{c\_array} \\ y_{\text{max\_length}} &= \max\left(\max\left(1, \sum_{i,j} w_{\text{ceil}}[i,j]\right)\right) \end{aligned} \tag{5.1}$$

This modification was added to the single-speaker Matcha-TTS, and the LJ Speech checkpoint[3] was used. For each duration step, the phrases "I heard them say peel", "I heard them say pill", "I heard them say fool", and "I heard them say full" were generated, resulting in 44 different stimuli (4 phrases $\times$ 11 manipulations).

5.2.2 Experimental procedure

We used the Gorilla Experiment Builder[4] to structure our experiments. The participants were first asked to provide demographic information on the language they first learned, their most commonly used daily language, their age and gender, and their self-rated English proficiency. We also collected pronunciation tests; however, this remains to be analyzed in the future.

Participants were then presented successive trials consisting of a single stimulus, for which they were asked to choose which of 2 alternative words ("pill" or "peel", or "full"

[3]urlhttps://drive.google.com/drive/folders/17C_gYgEHOxI5ZypcfE_k1piKCtyR0isJ

[4]https://gorilla.sc/

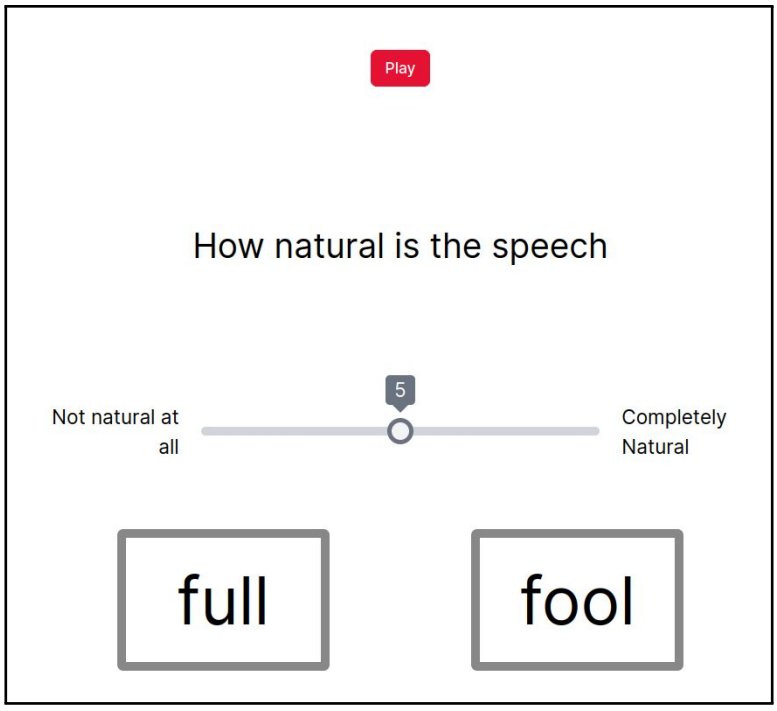

Figure 5.8: Screenshot of validation experiment set up in Gorilla.

or "fool") they thought it included (1-interval, 2-alternative forced choice). Each of the 44 stimuli (4 phrases, 11 manipulations) was presented 5 times, in random order, resulting in 220 trials. The three conditions, word and context, word only, and context only, were varied between-participants. There were two attention checks, and the experiment took roughly 30 minutes. In each trial, participants could listen to the phrase once. They were also asked to provide a naturalness MOS (nMOS) score from the MOS-X2 [148] at each trial. An example of the setup can be seen in Fig. 5.8.

Participants

We recruited N = 145 participants via Prolific. We maintained French as our target L2 population but recruited a smaller set of Mandarin and Japanese L1 speakers to confirm our reverse correlation results. There were N = 50 (17F, age = 34.36±9.74) English-L1 speakers (25 context + word, 25 context + word + noise), N = 75 (36F, age = 34.54±11.06) French-L1 speakers (25 context + word, 25 context only, 25 word only), N = 10 (7F, age = 28.70±7.82) Mandarin-L1 speakers (context + word), and N = 10 (8F, age = 36.40±13.44) Japanese-L1 speakers (context + word). The participant language demographics can be seen in Table 5.1.

Table 5.1: Participant language demographics for durational vowel control validation study

Category	**Count**
English-L1	
Daily Language	
English	50
Self-Rated English Proficiency 1(no proficiency) - 5(fluent)	
Score 4	4
Score 5	46
French-L1	
Daily Language	
French	52
English	21
French & English	2
Self-Rated English Proficiency 1(no proficiency) - 5(fluent)	
Score 1	2
Score 2	4
Score 3	11
Score 4	18
Score 5	40
Mandarin-L1	
Daily Language	
English	7
Mandarin	2
English & Mandarin	1
Self-Rated English Proficiency 1(no proficiency) - 5(fluent)	
Score 2	1
Score 3	2
Score 4	5
Score 5	2
Japanese-L1	
Daily Language	
Japanese	5
English	5
Self-Rated English Proficiency 1(no proficiency) - 5(fluent)	
Score 3	1
Score 4	2
Score 5	5

5.2.3 Results

Manipulation of both context and word

French-L1. Fig. 5.9 shows how different levels of manipulations of both context and word duration affect the normalized gain of accuracy (proportion of correct response) of French-L1 participants, relative to the accuracy obtained when no manipulation is applied. Baseline accuracy is relatively high, at 80.8% for "pill", 76.8% for "peel", 83.2% for "full" and 32.0% for "fool". As the French speakers had high English proficiency, it is unsurprising that participants perform relatively well on the baseline. Yet, we see a bias towards the lax vowel (left) (80.8% accuracy "pill" vs. 76.8% accuracy "peel"; 83.2% accuracy "full" vs 32.0% accuracy "fool") that we would like to overcome to improve comprehension. For "fool" (bottom right), specifically, the TTS struggled to synthesize clear formants, and even the EL1 speakers struggle to identify the correct word at the baseline (see *English-L1*).

We observe that we can overcome the bias towards lax vowels by increasing the length of tense vowels (right) ("peel": 1.6x/w-0.77x/c: $t(25)=2.57$, $p=.017$, 1.8x/w-0.71x/c: $t(25)=2.87$, $p=.008$, 2.0x/w-0.67x/c: $t(25)=3.36$, $p=.003$; "fool": 1.4x/w-0.83x/c: $t(25)=2.67$, $p=.013$, 1.6x/w-0.77x/c: $t(25)=4.38$, $p<.001$, 1.8x/w-0.71x/c: $t(25)=7.24$, $p<.001$, 2.0x/w-0.67x/c: $t(25)=5.28$, $p<.001$). We do not observe a significant improvement in performance for the lax vowels as the word becomes shorter. Still, we do see that, although the participants have high proficiency, they can be convinced to hear the tense vowel if the duration of a lax vowel becomes long (left) ("pill": 1.2x/w-0.9x/c: $t(25)=-2.75$, $p=.011$, 1.4x/w-0.83x/c: $t(25)=-2.30$,$p=.031$, 1.6x/w-0.77x/c: $t(25)=-3.32$, $p=.003$, 1.8x/w-0.71x/c: $t(25)=-3.29$, $p=.003$, 2.0x/w-0.67x/c: $t(25)=-4.11$, $p<.001$; "full": 1.2x/w-0.9x/c: $t(25)=-2.78$, $p=.010$, 1.6x/w-0.77x/c: $t(25)=-3.36$, $p=.003$, 1.8x/w-0.71x/c: $t(25)=-4.68$, $p<.001$, 2.0x/w-0.67x/c: $t(25)=-4.87$, $p<.001$).

English-L1. Fig. 5.10 shows how different levels of manipulations affect the gain in accuracy of English-L1 participants relative to baseline. Baseline accuracy is 96% for "pill", 94.4% for "peel" and "full", and, as above, 32% for "fool". We observe that contrary to the previous experiment where vowels were spectrally manipulated to be ambiguous, English speakers are now able to perform with nearly 100% accuracy for the "pill", "peel" and "full" and that, contrary to French-L1, they are mostly insensitive to the manipulation. It is possible that because they have more robust auditory representations of the spectral content of vowels in their native language, English speakers are able to use the formants of the vowel to differentiate the words and are not easily swayed by changes in duration. While this pattern of results seems to contradict the English-L1 kernels seen in the previous reverse-correlation experiment (Fig 5.6), we do, in fact, see evidence that even English-L1 participants may start relying on duration cues in situations where timbral cues are not easily processed. First, we observe that extreme shortening of "peel" does begin to affect the performance (top right) (0.55x/w-1.4x/c: $t(25)=-2.61$, $p=.015$, 0.5x/w-1.5x/c: $t(25)=-$

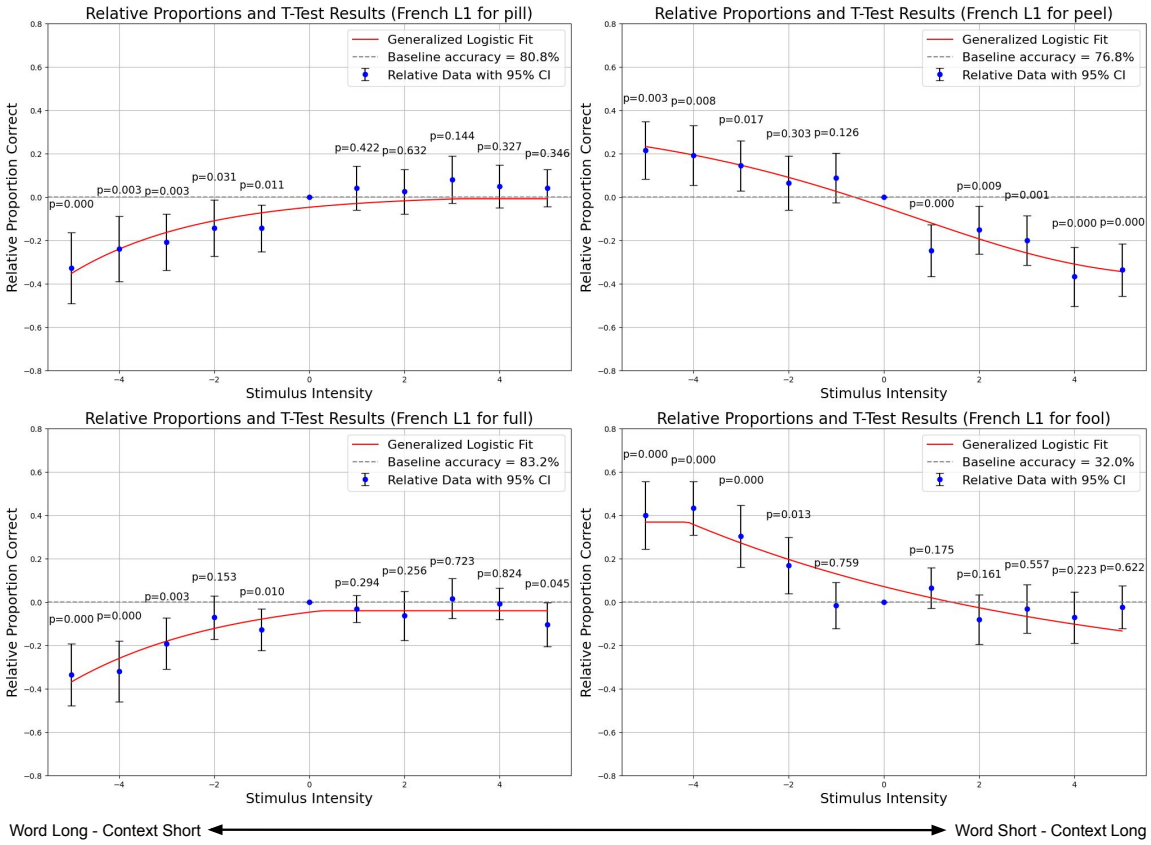

Figure 5.9: **Word identification performance in French-L1 participants, when both context and word are manipulated**: Proportion of correct responses for each level of duration multiplier, from 2.0x (left) to 0.5x (right) in the word and from 0.67x (left) to 1.5x (right) in the context, normalized relative to no manipulation (0.75x). P-values correspond to one-sample t-tests for the difference to the baseline. The red curve is a generalized logistic curve fit.

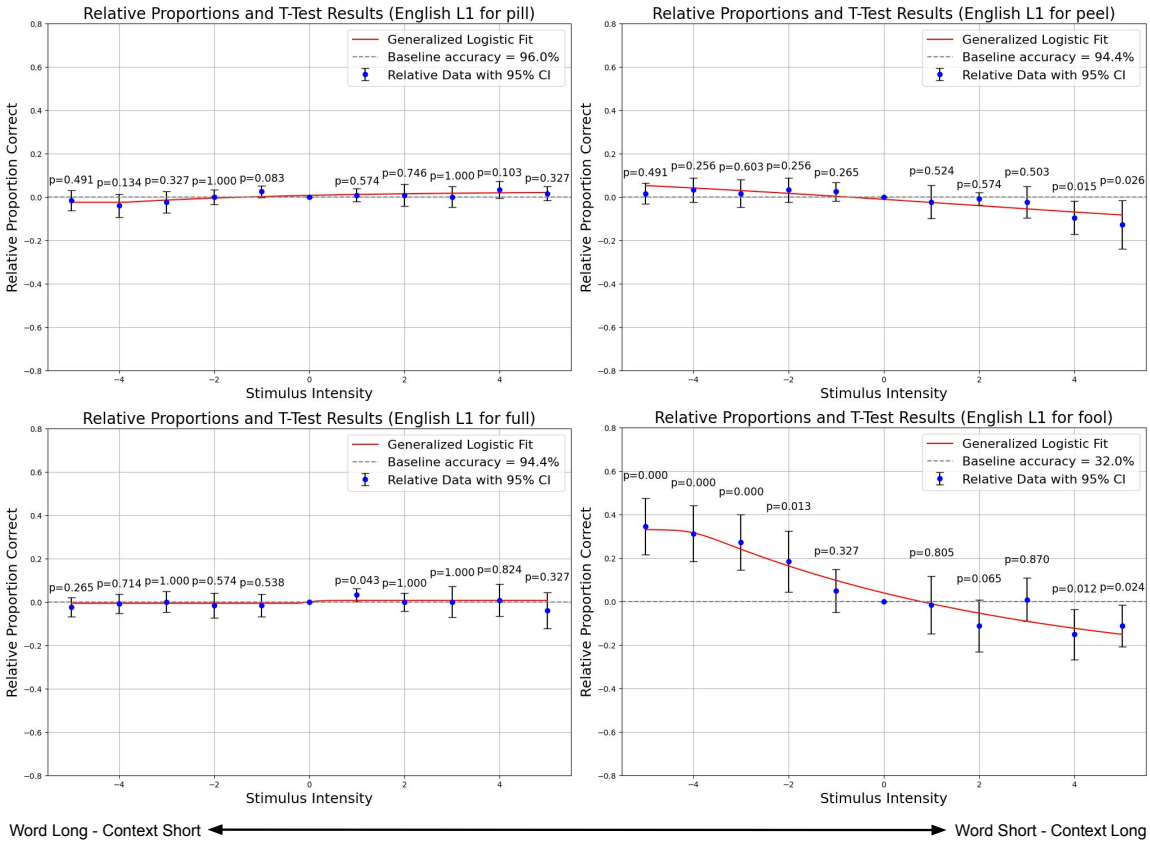

Figure 5.10: **Word identification performance in English-L1 participants, when both context and word are manipulated:** Proportion of correct responses for each level of duration multiplier, from 2.0x (left) to 0.5x (right) in the word and from 0.67x (left) to 1.5x (right) in the context, normalized relative to no manipulation (0.75x). P-values correspond to one-sample t-tests for the difference to the baseline. The red curve is a generalized logistic curve fit.

2.37, p=.026). Second, and most strikingly, when the TTS struggled to clearly generate the word "fool" (resulting in a much lower baseline performance), we see English speakers using the same duration cues as the French-L1 speakers, with performance increasing as the duration of "fool" increases (bottom right) (1.4x/w-0.83x/c: t(25)=2.70, p=.013, 1.6x/w-0.77x/c: t(25)=4.38, p<.001, 1.8x/w-0.71x/c: t(25)=5.01, p<.001, 2.0x/w-0.67x/c: t(25)=5.58, p<.001).

Mandarin-L1. Word identification performance in Mandarin-L1 participants can be seen in Fig. 5.11. Like FL1 speakers, ML1 speakers exhibited a slight bias towards lax vowels (left) (88.0% accuracy "pill" vs. 82.5% accuracy "peel"; 78.0% accuracy "full" vs 17.0% accuracy "fool"), albeit less statistically (note the smaller sample size). As in FL1 and EL1 for "fool", we observe an improvement in performance for tense vowels as the duration is lengthened (right). Interestingly, for "fool," the performance on the baseline was particularly

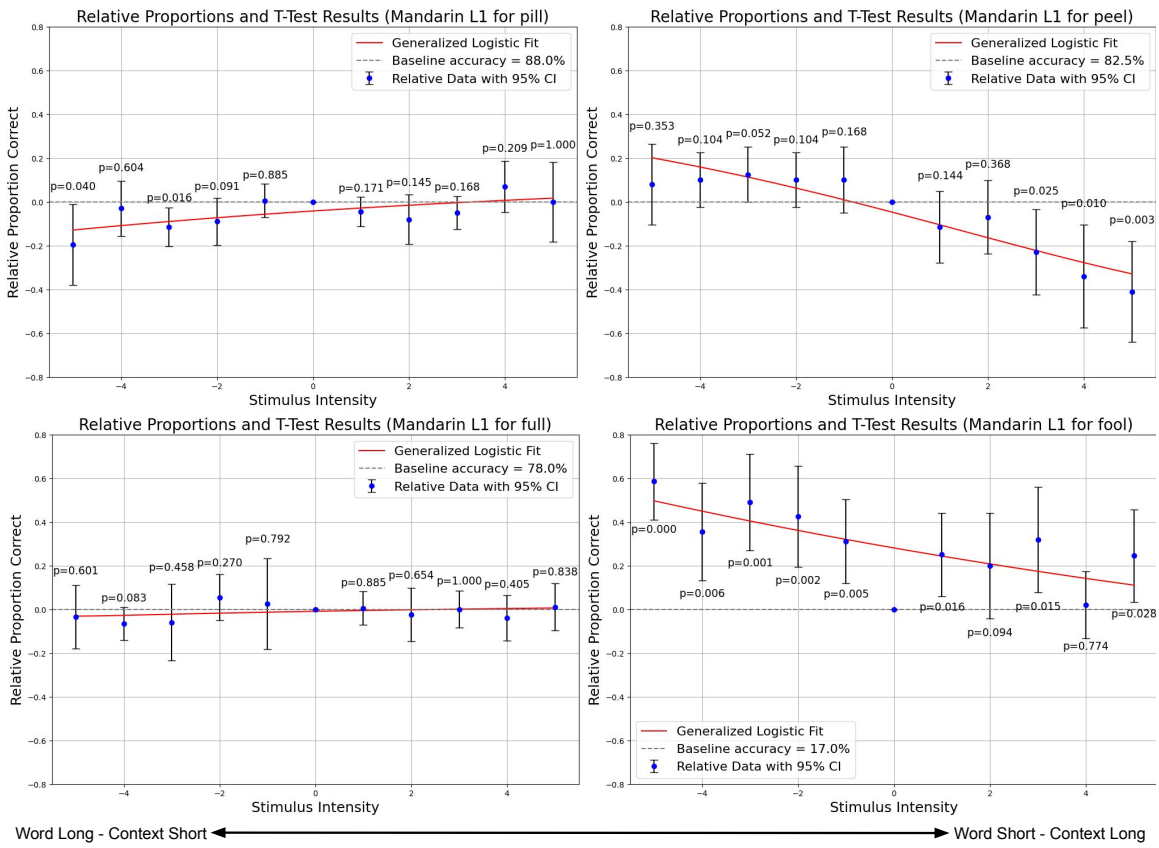

Figure 5.11: **Word identification performance in Mandarin-L1 participants, when both context and word are manipulated:** Proportion of correct responses for each level of duration multiplier, from 2.0x (left) to 0.5x (right) in the word and from 0.67x (left) to 1.5x (right) in the context, normalized relative to no manipulation (0.75x). P-values correspond to one-sample t-tests for the difference to the baseline. The red curve is a generalized logistic curve fit.

low, even compared to the shortened stimuli (bottom right); the rest of the curve displayed the same trend as we saw in FL1 and EL1.

Japanese-L1. Word identification performance in Japanese-L1 participants can be seen in Fig. 5.12. Despite reporting lower proficiency than the other L2 speakers, the Japanese speakers perform better than other L2 speakers on the baseline (98.0% accuracy for "pill", 88.0% for "peel"; 98.0% for "full"; 28.0% for "fool"), with a similar bias towards the lax vowel (left). Very similarly to French-L1 and, to some extent, to Mandarin-L1, Japanese speakers are sensitive to duration manipulations in the stimuli and can be easily swayed to hear the incorrect vowel when the duration is changed in the wrong direction, e.g. when "peel" becomes too short (top right) or "pill" becomes too long (top left) ("pill": 1.6x/w-0.77x/c: $t(10)=-2.54$, $p=.032$, 2.0x/w-0.67x/c: $t(10)=-2.45$, $p=.037$; "full": 1.2x/w-0.9x/c: $t(10)=-2.69$, $p=.025$, 1.4x/w-0.83x/c: $t(10)=-3.54$, $p=.006$, 1.6x/w-0.77x/c: $t(10)=-3.58$, $p=.006$, 1.8x/w-0.71x/c: $t(10)=-6.47$, $p<.001$, 2.0x/w-0.67x/c: $t(10)=-5.30$, $p<.001$,

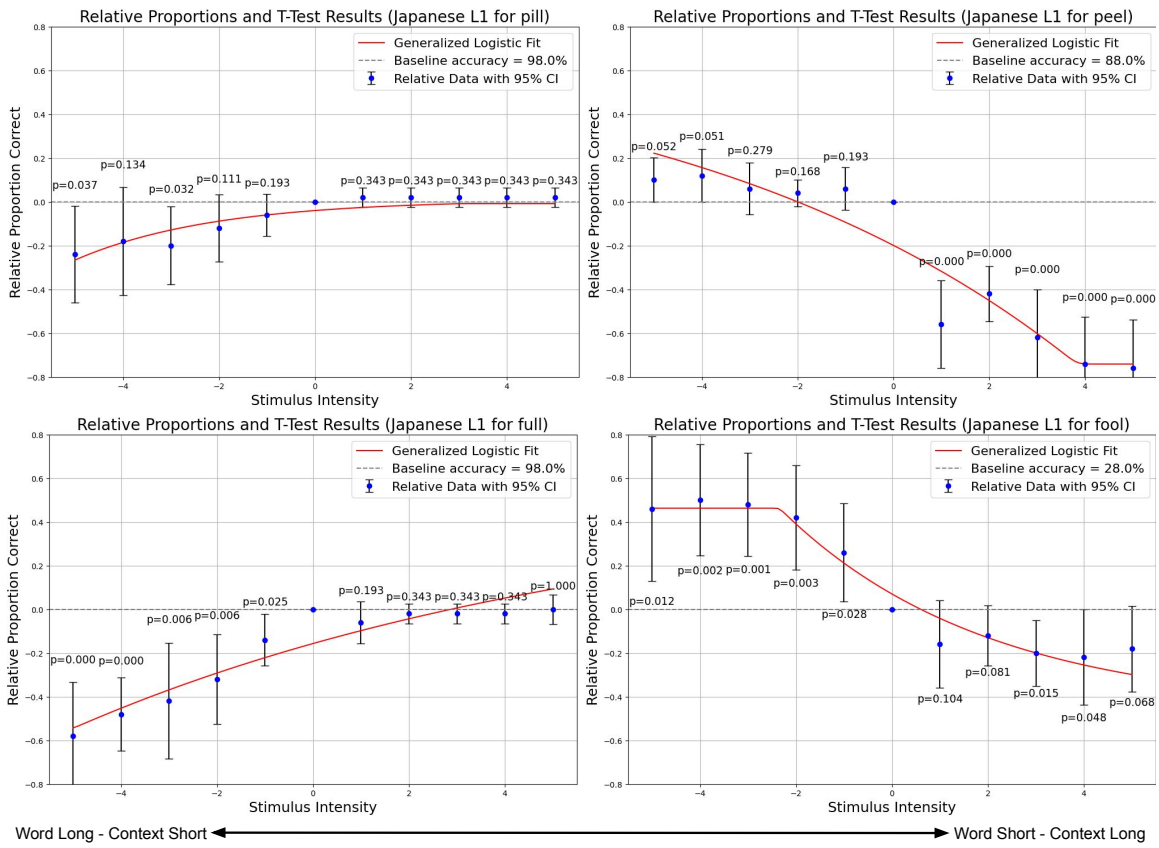

Figure 5.12: **Word identification performance in Japanese-L1 participants, when both context and word are manipulated:** Proportion of correct responses for each level of duration multiplier, from 2.0x (left) to 0.5x (right) in the word and from 0.67x (left) to 1.5x (right) in the context, normalized relative to no manipulation (0.75x). P-values correspond to one-sample t-tests for the difference to the baseline. The red curve is a generalized logistic curve fit.

"peel": 0.83x/w-1.1x/c: t(10)=-6.33, <.001, 0.71x/w-1.2x/c: t(10)=-7.58, p<.001, 0.63x/w-1.3x/c: t(10)=-6.43, p<.001, 0.55x/w-1.4x/c: t(10)=-7.83, p<.001, 0.5x/w-1.5x/c: t(10)=-7.75, p<.001). For "peel," we do not see a significant increase in accuracy by lengthening the word, which may be because of a ceiling effect (baseline was already at 88%). In "fool", as in all other language groups, we see a strong increase in performance with lengthening of the target word (bottom right) (1.2x/w-0.9x/c: t(10)=2.62, p=.028, 1.4x/w-0.83x/c: t(10)=3.99, p=.003, 1.6x/w-0.77x/c: t(10)=4.61, p=.001, 1.8x/w-0.71x/c: t(10)=4.44, p=.002, 2.0x/w-0.67x/c: t(10)=3.15, p=.012).

English-L1, with added noise. The previous results suggest that L2 speakers in 3 languages rely consequentially on duration cues for word identification. However, L1 speakers are already at ceiling performance based on spectral cues and may not need or even use duration cues for the task in typical conditions. However, English-L1 results for "fool" stimuli, a word which, for technical reasons, was found to have poor synthesis results with our

TTS system, suggest that, in more difficult or ambiguous listening conditions, even native speakers would default back to the same mechanism used as L2 speakers. To confirm this hypothesis, we wanted to explore if it is possible to mask formantic information with distortion and background noise and force L1 English speakers to behave like L2 for "peel" and "pill". "Fool" and "full" were not tested in this condition as we already observed the duration mechanism without the need for background noise.

We first attempted to overlay a loud background noise, yet through a pilot study, we found that this simply resulted in a loss of the target word, and participants resorted to random guessing. We then aimed to create a sound similar to loud-speakers in a metro station, i.e., a distorted and distanced sound, with a crowd in the background. To achieve this sound, we used Audacity[5]. First, a rectifier distortion was applied at 45%; then reverberation was added with a room size of 22%, a pre-delay of 10ms, a reverberance and dampening percentage of 50%, a tone low, tone high, and stereo width of 100%, and a wet and dry gain of -1bB. Finally, a background crowd noise was added so that all word duration in the phrase could be perceived, but it was difficult to recognize all of the speech clearly.

While baseline performance remains relatively high (84.5% "pill" and 81.8% "peel"), results, seen in Fig. 5.13, show an apparent effect of duration manipulations on word recognition performance both for "pill" and 'peel", which mirrors what is seen in L2 speakers. These results, therefore, confirm that the duration perception mechanism evidenced in reverse-correlation experiments is one used by both L1 and L2 speakers (with French, Mandarin and Japanese L1s) of English as a fall-back strategy when spectral information on the word itself is unreliable, both for internal (L2 speakers) or external (L1 with noise) reasons.

Manipulation of context-only and word-only

Finally, after confirming reverse-correlation results for simultaneous context and word manipulations of duration, we explored the contribution of manipulating only the context or only the word (i.e. applying either 0.67x to 1.5x in the context and 1x in the word or 2.0x to 0.5x in the word and 1x in the context). We focused only on L1 French speakers, as the results above showed the effect could be seen identically in the other populations.

We observed that although reverse correlation kernels showed effects of duration both outside and inside the word, word identification performance was more strongly driven by word duration than context duration. The context manipulations resulted in only very slight improvements over the baseline (Fig. 5.14), and modifying the word had as strong of an effect as modifying both the context and the word simultaneously (Fig. 5.9).

[5]https://www.audacityteam.org/

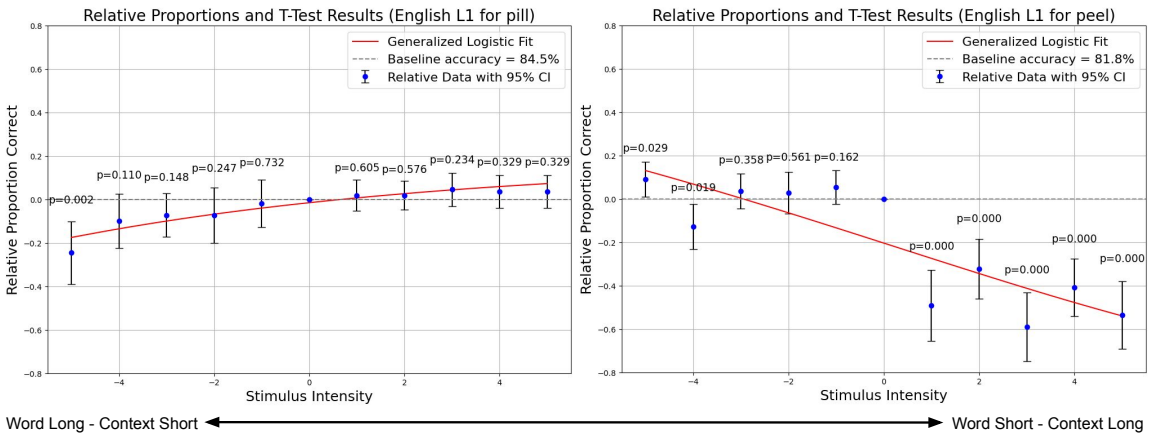

Figure 5.13: **Word identification performance in English-L1 participants, when both context and word are manipulated, in the addition of background noise and distortion.** Proportion of correct responses for each level of duration multiplier, from 2.0x (left) to 0.5x (right) in the word and from 0.67x (left) to 1.5x (right) in the context, normalized relative to no manipulation (0.75x). P-values correspond to one-sample t-tests for the difference to the baseline. The red curve is a generalized logistic curve fit.

5.2.4 Discussion

Reverse correlation experiments in Section 5.1 showed that vowel perception was driven psychophysically by duration cues both in the target word and in its preceding context, but the question as to whether this mechanism could be used causally to drive word identification performance in ecological situations remained open. The present results showed that this was indeed the case, using non-ambiguous vowels and expanding the results to another tense/lax vowel pair (‘full” (/fʊl/) and “fool” (/ful/).

The mechanism was most strongly observed in the case of tense vowels, where there is generally a lower level of perception accuracy on the baseline. For these vowels, applying a lengthening allowed for an improvement in word identification, and a shortening of tense vowels decreased performance. In comparison, little advantage was seen when shortening lax vowels. This may be due to the bias towards lax vowels already existing in the baseline or to the fact that the words may become too short and difficult to hear. Instead, we observed that it is important not to lengthen words containing lax vowels, as this can lead to confusion with their tense counterparts.

Strikingly, this was confirmed with remarkable consistency for all non-native FL1, ML1, and JL1 speakers. This is especially remarkable given the vast difference in reliance on duration cues for vowel differentiation in these three languages, i.e. Japanese relies primarily on duration, Mandarin relies primarily on pitch but uses duration as a secondary cue, and French is nearly void of a reliance on duration cues (see Sec. 5.1). In EL1 speakers, eliciting this mechanism was more difficult, but we were able to evidence it when spectral cues were either poorly synthesized (“fool”) or masked with distortion and noise. This pattern of

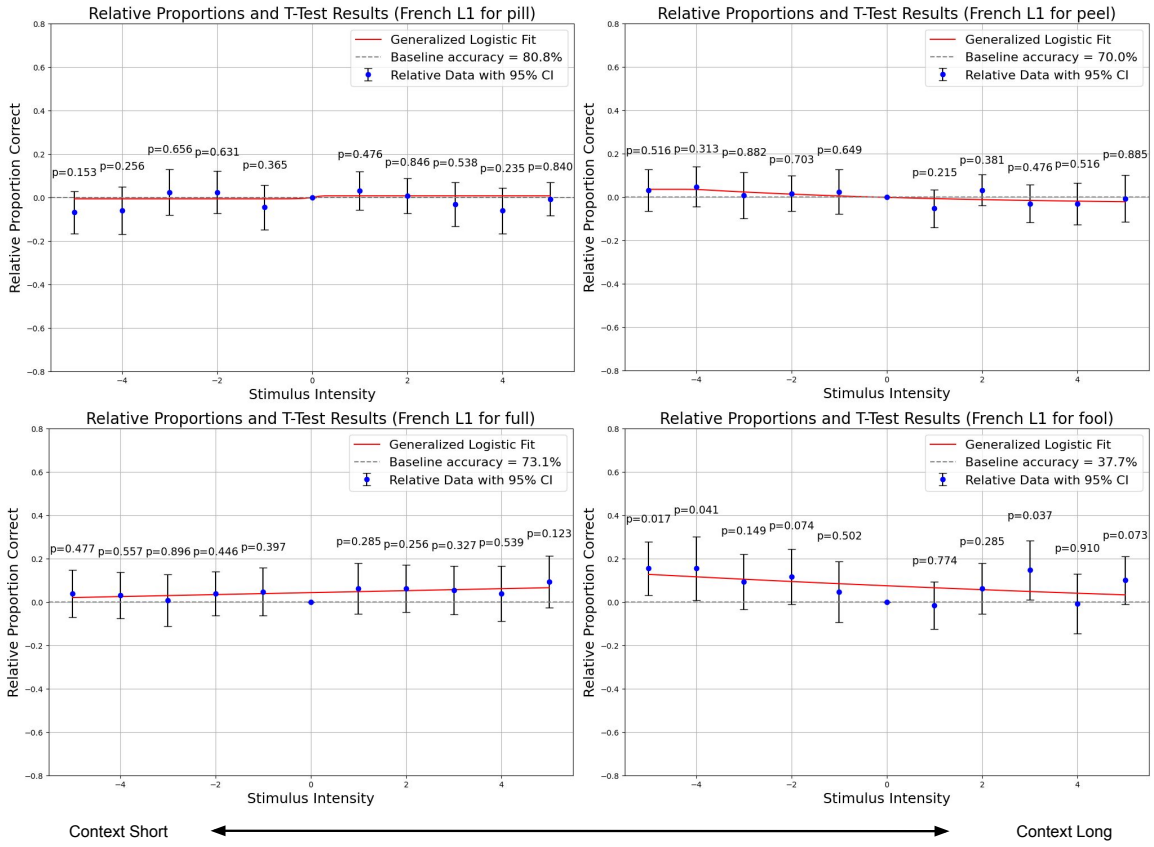

Figure 5.14: **Word identification performance in French-L1 participants, when both only the context is manipulated:** Proportion of correct responses for each level of duration multiplier from 0.67x (left) to 1.5x (right) in the context, normalized relative to no manipulation (0.75x). P-values correspond to one-sample t-tests for the difference to the baseline. The red curve is a generalized logistic curve fit.

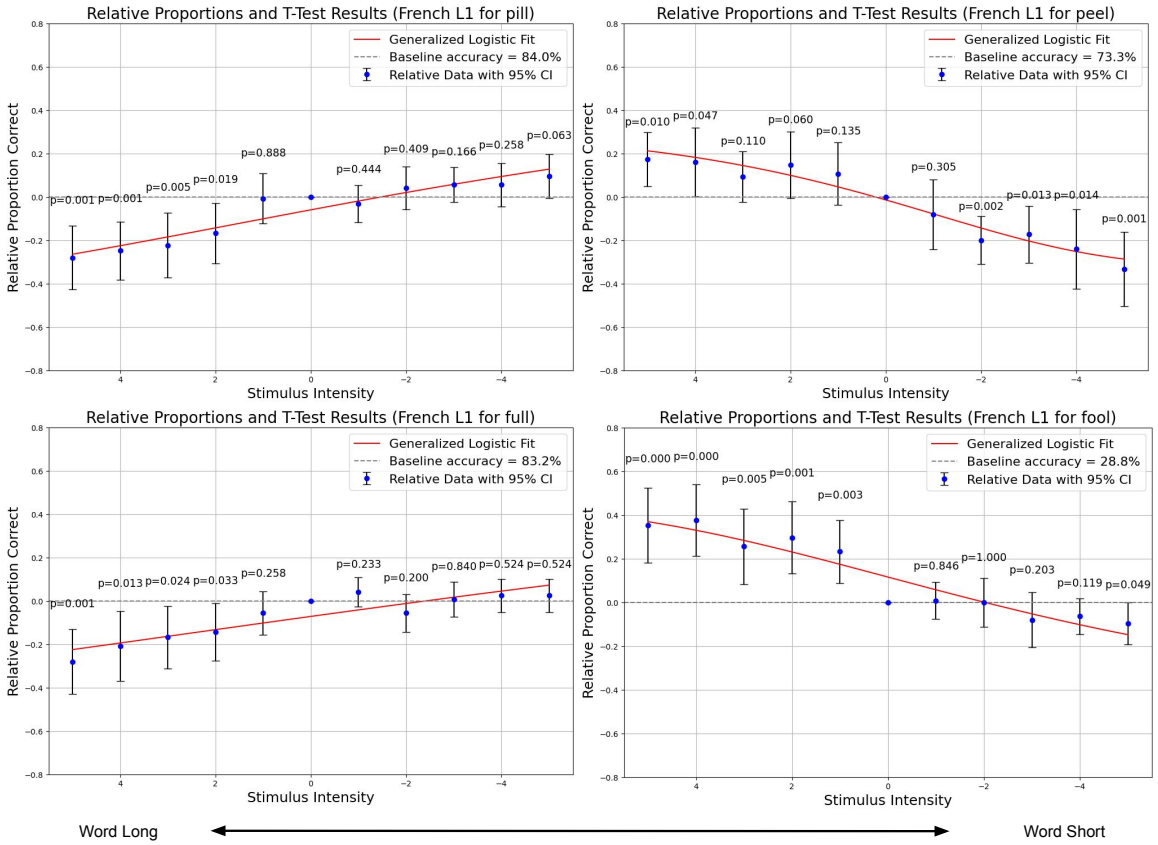

Figure 5.15: **Word identification performance in French-L1 participants, when both only the word is manipulated:** Proportion of correct responses for each level of duration multiplier from 2.0x (left) to 0.5x (right) in the word, normalized relative to no manipulation (0.75x). P-values correspond to one-sample t-tests for the difference to the baseline. The red curve is a generalized logistic curve fit.

result is consistent with a universal duration-perception mechanism which acts as a fall-back strategy, which EL1 speakers who can easily use the formants to differentiate minimal pairs will only use in challenging listening conditions but may be used routinely, and perhaps even principally, by L2 speakers.

Although we collected MOS scores for naturalness and do not report on their detailed analysis here, we saw no significant differences in naturalness ratings from the baseline as the duration changes were applied. We also did not see differences in naturalness ratings between context and word, context-only and word-only. However, we do note that EL1 speakers began to trend towards lower naturalness in both extremes of stretch and compression and had a lower standard deviation across their responses overall. It appears that L2 speakers have a wide range of expectations and acceptance of naturalness that may relate to their English proficiency. This remains future work to explore.

Taken together, these results provide a promising strategy to improve the clarity of synthesized speech for L2 speakers by manipulating duration cues to enforce the correct perception of tense/lax alternatives. In more detail, as we did not see an improvement in either performance or naturalness through the addition of duration changes in the context, we hypothesize that a word-only modification should be sufficient to improve comprehension for L2 listeners. Moreover, as the duration effects seen above were asymmetric, we make the hypothesis that lax vowels should remain at the base speaking rate, but words containing a tense vowel that can be easily confused with a lax vowel minimal pair should be lengthened relative to the rest of the phrase. In the following, we implement this strategy in a complete TTS system, using a parsing technique to automatically identify portions of the phrase that should benefit from durational changes and validate that this strategy improves speech comprehension compared to two other control strategies (slowing down the difficult word or slowing down the whole sentence).

5.3 L2 Clarity TTS

Once we had confirmed we could elicit behavioural changes to improve the perception of tense/lax vowels using simple linguistically driven duration changes, we added the changes we had applied manually as a new "L2 clarity" mode for Matcha-TTS in [79]. Moreover, we add the last tense/lax vowel pair: /ɑ/ (cot) and /ʌ/ (cut).

5.3.1 Pilot clarity mode

Initially, to enable L2 clarity mode in Matcha-TTS, we added a clarity flag that could be set to "True" or "False" at synthesis. If set to true, it automatically applied a stretch to words containing a tense vowel to increase clarity. This, however, becomes complicated when there are multiple vowels in a word, and the speech rate of a phrase can quickly become awkward with too many stretched words. We decided to eliminate context changes to minimize the

number of duration changes in a phrase as we previously saw that the perception was driven primarily by the target word, and as we discovered through our validation, a lax vowel's duration does not need to be modified to increase clarity.

In order to reduce unnecessary and frequent duration changes we created a list of function words such as "be" and "he" that were ignored for clarity treatment. We ensured that no function words in this "ignore" list had a minimal pair, such as "been" (/bin/), which can be confused with "bin" (/bɪn/). Moreover, diphthongs and words ending in /i/ were ignored. The model first parsed the phrase through several steps:

1. Parse each word to see if it contains a tense or lax vowel
2. If the word is in the list of function words, it is ignored
3. If the word contains tense vowels but no lax vowels, the clarity modification is applied
4. If the word contains both tense and lax vowels, if the tense vowel has primary stress, the clarity modification is applied

The clarity modification was applied as in Algorithm 5.1. A 1.6x stretch was applied across the entire word with a gradual ramp up and down to the base speech rate over the 6 phonemized items (if there are at least 6 phonemized items between target words, otherwise only as many as are available) preceding and following the target word. Six items were chosen as this encompasses two phonemes preceding and following the target word (approximately 200-300ms [178] as per our reverse correlation results). We also apply a 1x stretch (the base speech rate) to lax words simultaneously, following the same parsing above to ensure that lax vowels are not stretched by surrounding tense vowel words. A 1.6x stretch was chosen to minimize duration changes, because as in our validation, we saw that perception performance quickly improved from the baseline by increasing the duration of the target word, and there was minimal improvement in performance between a 1.6x and 2.0x stretch.

We, however, discovered through a pilot study of French-L1-English-L2 speakers (N=50) that although we attempted to control unnecessary changes in duration, there still remained many duration changes in the context leading up to and following target words. Because of this, our results showed little improvement over the baseline, and the overall results were inconsistent. Interestingly, although in our validation study (Sec. 5.2) we found that the context changes alone were not enough to influence perception, these pilot results echo our reverse correlation findings that we need a contrasting and consistent difference in duration between the context and the target word to induce improvement in tense vowel comprehension, i.e., having too many duration changes in the phrase does not create a strong enough contrast in the target word to illicit improvements in comprehension. As such, we needed to control the duration changes in the TTS in a more defined manner.

5.3.2 Final L2 clarity mode

Based on our findings in the pilot experiment, we decided to include a markup to control which words are emphasized. As the text for a TTS must be generated either by a human or a large language model (LLM), these markups can be added to words that are either perceived as important, difficult, or lacking context in the sentence and should be emphasized by the TTS at the time of the text generation. This markup means that we no longer need to ignore function words, nor parse the entire sentence, instead, only the flagged words are parsed for clarity treatment. Once the clarity flag is enabled, as in the pilot, the user then surrounds difficult words with exclamation points, e.g., "!peel!", allowing the TTS to parse the only words to be treated for clarity. The parsing and duration modifications were then applied as follows:

1. Parse each flagged word to see if it contains a tense or lax vowel
2. If the word contains tense vowels but no lax vowels, the clarity modification is applied
3. If the word contains both tense and lax vowels, if the tense vowel has primary stress, the clarity modification is applied

5.3.3 Stimulus generation

We tested 4 different TTS styles: 1. **Base**: the base Matcha-TTS (0.75x speech rate), 2. **Stretch**: 1.2x (0.75×1.6) speech rate applied across the entire phrase, 3. **Emphasis**: 1.6x stretch applied across all target words (0.75x speech rate elsewhere), and 4. **Clarity**: 1.6x stretch applied across the target words containing a tense vowel (0.75x speech rate elsewhere).

In our reverse correlation and validation experiments, only a **single target word**, always at the end of a phrase, was tested. To test the robustness of the L2 clarity TTS when the target words occurred in other parts of the phrase, we tested clarity mode in several contexts. First, single target word phrases where the target word was in the middle of the phrase were tested. We then tested **double target word** phrases (e.g., "Write down dull and doll"), where the targets could be at the end of the phrase, in the middle of the phrase, or at the beginning of the phrase. Additionally, we explored the performance when the TTS had a combination of tense and lax vowels in a single phrase; we tested a tense and a lax minimal pair, a tense and a lax that are not minimal pairs, two tense and two lax vowels. We aimed to use a variety of starting and ending consonants surrounding the target vowels, and all words had a minimal pair in English. We tested all tense/lax vowel pairs: /i/ (peel) and /ɪ/ (pill), /u/ (fool) and /ʊ/ (full), /ɑ/ (cot) and /ʌ/ (cut). Lastly, we aimed not to semantically bias phrase meanings towards any word, that is, given the context outside of the target word neither word in the minimal pair made more logical or semantic sense. To do this, we asked ChatGPT-4o to provide a starting list of neutral phrases with varying

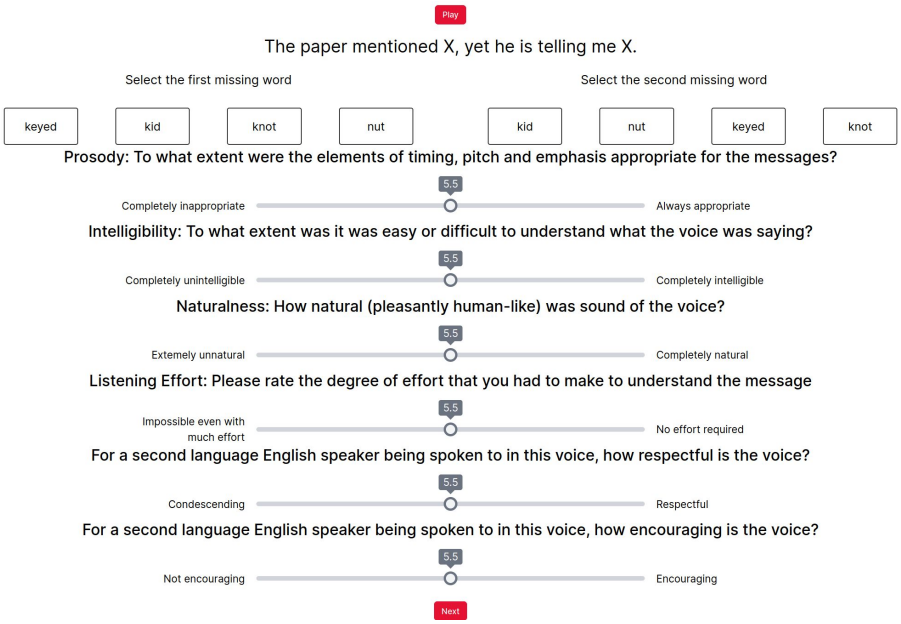

Figure 5.16: Screenshot of L2-TTS experiment set up in Gorilla.

lengths. These were then selected and modified to create our final list of 16 phrases and insert our target words (Appendix C).

Because participants would hear each of the phrases 4 times (each TTS style had the same phrases), we added 1 confusion phrase (randomly assigning a TTS style) for each of the phrases. A confusion phrase was the same phrase context using the opposite minimal pair, for example: "She kept mentioning cot during the conversation." and the confusion phrase "She kept mentioning cut during the conversation." Lastly, for the single target word TTS, "clarity" TTS was the same as "emphasis" for a single tense vowel target and "base" for a single lax vowel target. As such, we assessed the TTS in terms of the length of the target words rather than TTS styles and ensured no repeated phrases with the same treatment.

5.3.4 Experimental procedure

We conducted an experiment to assess the objective (through word error rate) and subjective (through mean opinion scores) performance of French L1 English L2 listeners using our "clarity mode" compared to the baseline models (i.e., "base", "stretch", "emphasis"). In an online experiment, participants chose to receive instructions in either English or French. They then provided demographic information on the language they first learned, their most commonly used daily language, their age and gender, and their self-rated English proficiency. Each participant was randomly assigned to start in either the single- or double-word trial.

The participants were shown the phrase with the target words removed, e.g., "Write down the word X followed by the word X on the paper," and could play the audio only once. They then selected which words they heard and were told (in the double-word case)

that it was possible to hear the same word twice. It was never the case that they would hear the same word, however, this instruction was added to encourage participants not to base their choice for the second word on what they believe they heard in the first word. The missing word was selected from a list of 4 words: for the single-word trial, 2 words were the tense/lax vowel minimal pair (e.g., beat, bit); the third word was a minimal pair word with another vowel (e.g., bat); and the final word was dissimilar from the other three choices (e.g., shop). The dissimilar choice functioned as an attention check. In the double word case, when the two words were a minimal pair, the selection for the remaining two words was as per the single-word trial; if the two words were not a minimal pair, the 4 choices were the two words and their minimal pairs. The order of the phrases was randomized for each participant. The experiment set up can be seen in Fig. 5.16.

Once the participants selected which word they heard, they responded to a questionnaire containing the MOS-X2 [148] naturalness (nMOS), intelligibility (iMOS), and prosody (pMOS) scores, as well as an intelligibility question for listening effort (eMOS) ("Please rate the degree of effort you had to make to understand the message") from MOS-X [148]. The listening effort question was added to understand if there is a difference between being able to understand the phrase and how much effort was required to understand the phrase. They then responded to two questions from L2-directed speech research [69]: "For an English second language speaker being spoken to with this voice, how respectful is the voice?" (Resp.: condescending - respectful), "For an English second language speaker being spoken to with this voice, how encouraging is the voice?" (Enc: not encouraging - encouraging). These questions were to help us understand if slowing down or adding emphasis makes the TTS sound condescending, as can be the case with L2-directed speech [68, 69]. All scores were on a 10-point Likert scale.

Participants

We recruited N=56 participants (29F, age = 34.59±10.67) via Prolific. Once again we maintained French as our target L2 population and the participant language demographics can be seen in Table 5.2. Additionally, 71.4% of participants chose to have the instructions in French.

5.3.5 Results

Single word

To confirm our previous findings, we expect improved performance in WER with lengthened tense vowels but decreased performance for the same treatment in lax vowels. We computed one-way ANOVAs (Type II) followed by a post-hoc Tukey HSD as well as word error rates (WER) as the proportion of incorrect target words for each type of target word treatment:

Table 5.2: Participant language demographics for L2-TTS study

Category	Count
Daily Language	
French	40
English	14
French & English	2
Self-Rated English Proficiency 1(no proficiency) - 5(fluent)	
Score 2	2
Score 3	5
Score 4	20
Score 5	29

"base" has the baseline duration on the target, "target stretch" is 1.6x the duration of the baseline on the target, and "full stretch" is 1.2x the duration on the entire phrase.

We observed that, through our "clarity" mode stretch applied to tense-vowel-containing words, we could overcome the bias towards lax vowels, i.e., the fact that participants were more likely to respond, for example, "pill" than "peel," which we observed in our validation study. By lengthening the target tense-vowel-containing word, the WER on the baseline ("base tense") was reduced from 60.23% to 29.48% for "target stretch tense" (Table 5.3, Fig. 5.17). Importantly, we also confirmed that, indeed, stretching lax vowels, as is typical in L2-directed speech, results in more errors (28.90% vs. 16.86%). Lastly, the "stretch" TTS resulted in a slight reduction in WER over the baseline for tense vowels, yet not as much as for only emphasizing the target word. This suggests that a difference between the target word and its context was important for determining vowel perception [78].

Surprisingly, despite the objective improvements seen in the WER, the participants perceived a "target stretch" target word as significantly more intelligible for lax vowels (Tables 5.3 and 5.4, Figs. 5.18, 5.19). Moreover, "target stretch" also required less listening effort (although not significantly). Furthermore, although the "full stretch" TTS shows improvements in WER over the baseline for tense vowels, it was found to be significantly less respectful and encouraging in all cases. We also saw that the "full stretch" TTS was considered to be less natural and had worse prosody than the other TTS styles. This confirms our expectation that maintaining a natural (for L1 speakers) speech rate outside of difficult words can make the L2 listeners feel less patronized by the voice. Interestingly, the "target stretch" lax was rated as more natural and as having better prosody than "base" lax vowel, once again, despite have lower objective comprehension scores. These findings are particularly interesting as they suggest that L2 speakers may be overconfident in their ability to use the formant cues when the vowel is lengthened.

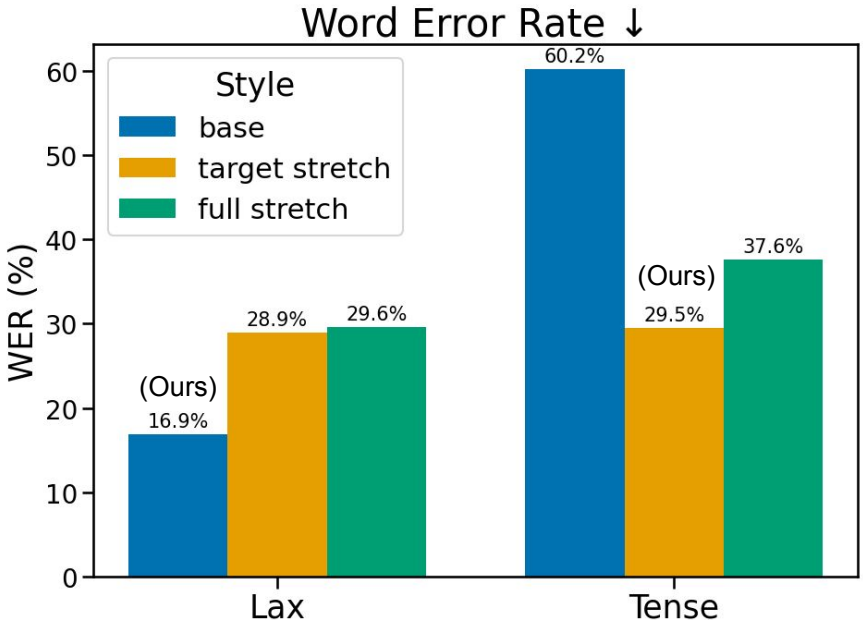

Figure 5.17: Single target word results: our clarity mode has the lowest WER for lax-base (left) and tense-target stretch (right).

Table 5.3: Single target word results: word error rate, intelligibility (iMOS), naturalness (nMOS), effort (eMOS), prosody (pMOS), encouragement (Enc) and respect (Resp.) Our clarity mode, using base lax and target stretch tense, achieves the lowest WER.

TTS Style	**WER ↓**	**iMOS ↑**	**nMOS ↑**	**eMOS ↑**	**pMOS ↑**	**Enc. ↑**	**Resp. ↑**
Base lax (Ours)	**16.86**%	7.63	7.11/***	7.33	7.20/***	6.44/***	6.63/**
Target stretch lax	28.90%	**8.18***/**	**7.91****/***	**7.73**	**7.87***/***	**7.03***/***	**7.32***/***
Full stretch lax	29.65%	7.53	4.18	7.40	5.83	5.54	5.76
Base tense	60.23%	7.57	7.07/***	7.18	7.22/***	6.45/***	6.53/**
Target stretch tense (Ours)	**29.48**%	**8.18***/	**7.71**/***	**7.69**	**7.80**/***	**6.84**/***	**7.00**/***
Full stretch tense	37.57%	7.68	4.27	7.61	6.13	5.49	5.68

Significance values are presented as: {significance over long or short TTS}/{significance over stretch TTS}

Table 5.4: Single target word results: ANOVA and Tukey test statistics on MOS Likert scores.

TTS Style	**F-Statistic (5,1029)**	**P-Value**	**Significant Tukey Results**
iMOS	5.12	<.001	Lax: targetS/base,fullS p=.043, p=.008; Tense: targetS/base p=.016
nMOS	110.83	<.001	Lax: all/fullS p<.001 targetS/base p=.006; Tense: all/fullS p<.001
eMOS	2.21	0.051	N/A
pMOS	30.24	<.001	Lax: all/fullS p<.001 targetS/base p=.028; Tense: all/fullS p <.001
Enc.	21.45	<.001	Lax: all/fullS p<.001 targetS/base p=.036; Tense: all/Stretch p<.001
Resp.	17.90	<.001	Lax: targetS/fullS,base p<.001, p=0.022 base/fullS p=0.001; targetS/fullS p<.001 base/fullS p=0.001

Tukey results are presented as: {higher value}/{lower value}
fullS = full stretch, targetS = target stretch

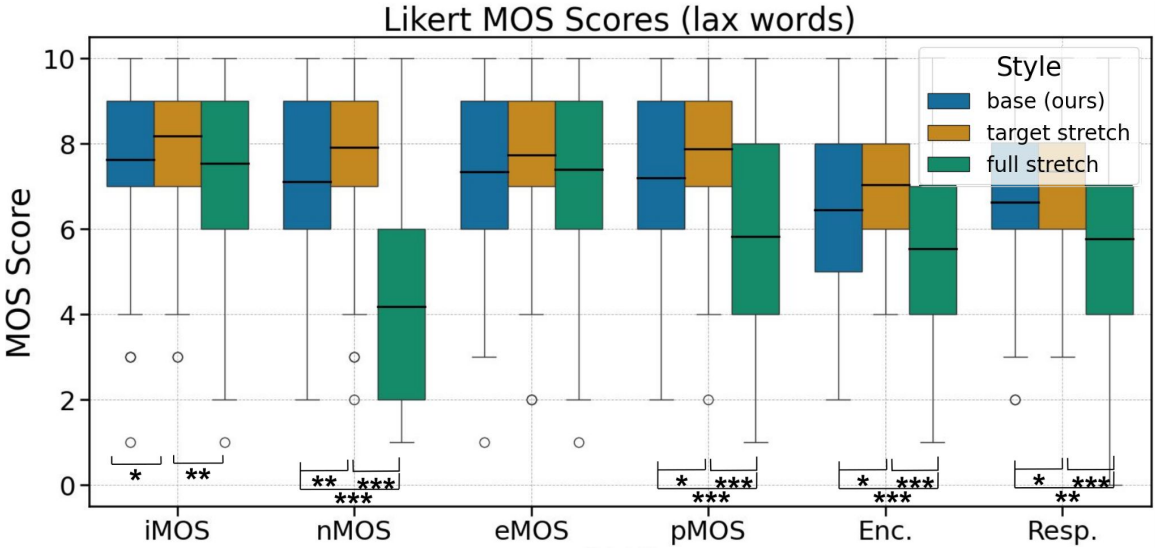

Figure 5.18: Single target word results: Likert MOS scores of lax vowel containing words for intelligibility (iMOS), naturalness (nMOS), effort (eMOS), prosody (pMOS), encouragement (Enc) and respect (Resp.) for base (ours), target stretch, and full stretch. Stretching the target word has significantly higher perceived intelligibility than both full stretch as the baseline, despite having lower objective comprehension performance. Both the baseline and target word stretch has significantly higher naturalness, prosody, encouragement and respectfulness over full stretch, and target word stretch over the baseline.

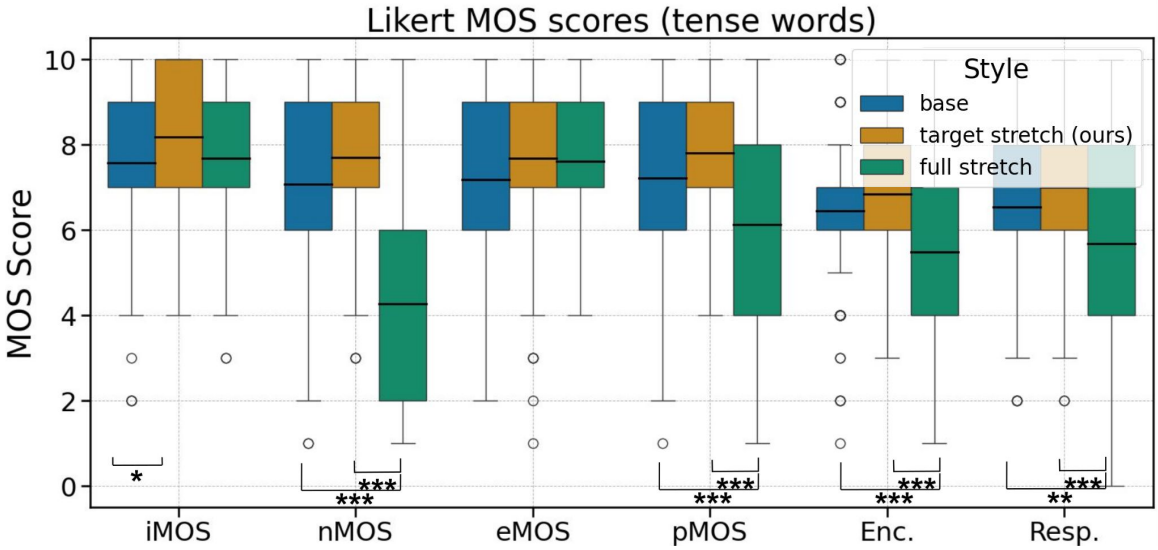

Figure 5.19: Single target word results: Likert MOS scores of tense vowel containing words for intelligibility (iMOS), naturalness (nMOS), effort (eMOS), prosody (pMOS), encouragement (Enc) and respect (Resp.) for base, target stretch (ours) and full stretch. Stretching the target word has significantly higher perceived intelligibility over the baseline. Both the baseline and stretching the target word have significantly higher naturalness, prosody, encouragement and respectfulness than the full stretch.

Table 5.5: Double target word results: total word error rate, tense word error rate, and lax word error rate.

TTS Style	**WER ↓**	**tWER ↓**	**lWER ↓**
Base	24.30%	30.00%	18.66%
Stretch	19.82%	17.99%	21.65%
Emphasis	24.44%	20.06%	28.82%
Clarity (Ours)	**15.15**%	**14.38**%	**15.92**%

Double word

Once again, we computed WER as the proportion of correct target words for each TTS style. We also assessed differences between tense (tWER) and lax (lWER) vowels.

We observed the lowest WER with our proposed "clarity" TTS, both overall and for lax vowels, while vastly improving the performance for tense vowels over the baseline (relative 50% improvement) (Table 5.5, Fig. 5.20). Again, the "base" TTS showed a bias towards lax vowels (tWER: 30.00%, lWER: 18.66%), that we were able to overcome with our "clarity" mode. Additionally, we see that simply emphasizing all difficult words that had a tense/lax minimal pair decreases performance in lax vowels. Lastly, we observed "stretch" having improved performance over both "emphasis" and "base" for tense vowels. Yet, as in the "emphasis" condition, we saw participants struggle to understand the short, lax vowels. Although the tense vowel words between "emphasis" and "clarity" and lax vowel words between "clarity" and base had the same duration, "clarity" mode still had a lower WER when specifically comparing these words. This likely resulted from the phrases containing both a tense and a lax vowel, where the L2 participants could use duration differences between the two words to more easily differentiate the words. A more in-depth exploration of these differences remains a topic for future work.

We once again observed the fascinating result that the participants rate the "emphasis" TTS the highest in all categories (although in this case, the scores are not significantly higher than those for the "clarity" TTS) (Tables 5.6 and 5.7, and Fig. 5.21), despite the WER being higher for this TTS than both "clarity" and "stretch" TTS styles. We observed that "stretch" is less natural and has poorer prosody than all other TTS styles, despite showing objective improvements in WER over "emphasis" and "base". Lastly, we found that L2 participants rated both speaking too fast ("base") and too slow ("stretch") as being less respectful and less encouraging.

Whisper ASR

Speech synthesis studies often use human MOS scores but rely on ASR for transcription accuracy, as it correlates well with L1 intelligibility [179]. In this section, we explore how ASR relates to L2 performance and whether it uses the same duration cues as L2 participants.

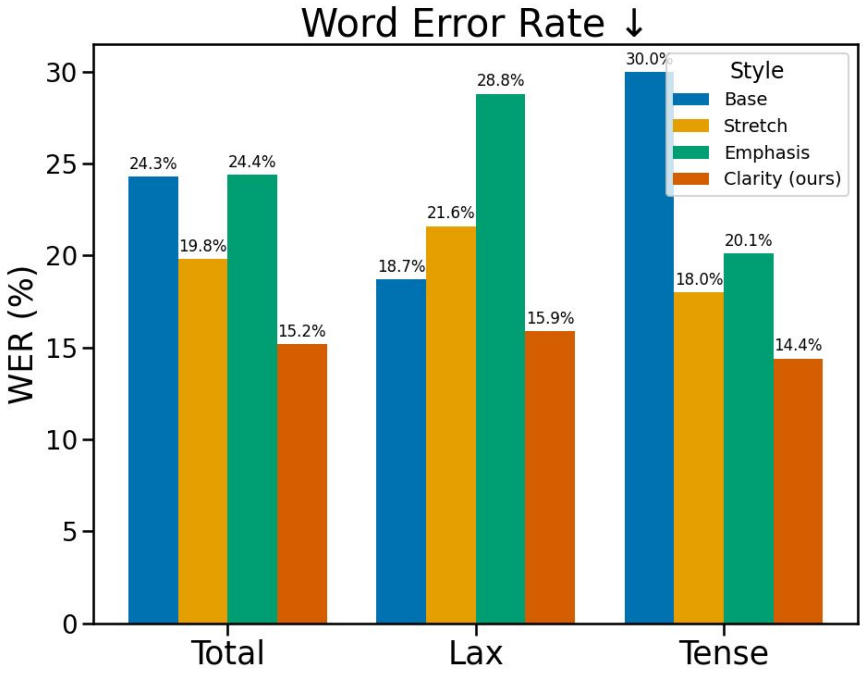

Figure 5.20: Double target word results: our clarity mode has the lowest total WER (left), lax WER (center) and tense WER (right).

Table 5.6: Double target word results: intelligibility, naturalness, effort, prosody, encouragement and respect.

TTS Style	**iMOS** ↑	**nMOS** ↑	**eMOS** ↑	**pMOS** ↑	**Enc.** ↑	**Resp.** ↑
Base	7.30	6.95/***	6.60	7.16/***	6.25	6.70
Stretch	7.94***/	4.93	**7.53*****/	6.40	6.02	6.46
Emphasis	**8.06*****/	**7.71*****/***	7.43***/	**7.98*****/***	**6.97*****/***	**7.42*****/***
Clarity (Ours)	7.83***/	7.54***/***	7.25***/	7.77***/***	6.72***/***	7.16***/***

Significance values are presented as: {significance over base TTS}/{significance over stretch TTS}

Table 5.7: Double target word results: ANOVA and Tukey test statistics on MOS Likert scores.

TTS Style	**F-Statistic (3,2504)**	**P-Value**	**Significant Tukey Results**
iMOS	20.15	<.001	all/Base p<.001
nMOS	20.15	<.001	all/Stretch p<.001; Emphasis,Clarity/Base p<.001
eMOS	24.15	<.001	all/Base p<.001
pMOS	77.15	<.001	all/Stretch <.001; Emphasis,Clarity/Base p<.001
Enc.	35.74	<.001	Emphasis,Clarity/Stretch,Base p<.001
Resp.	35.87	<.001	Emphasis,Clarity/Stretch, Base p<.001

Tukey results are presented as: {higher value}/{lower value}

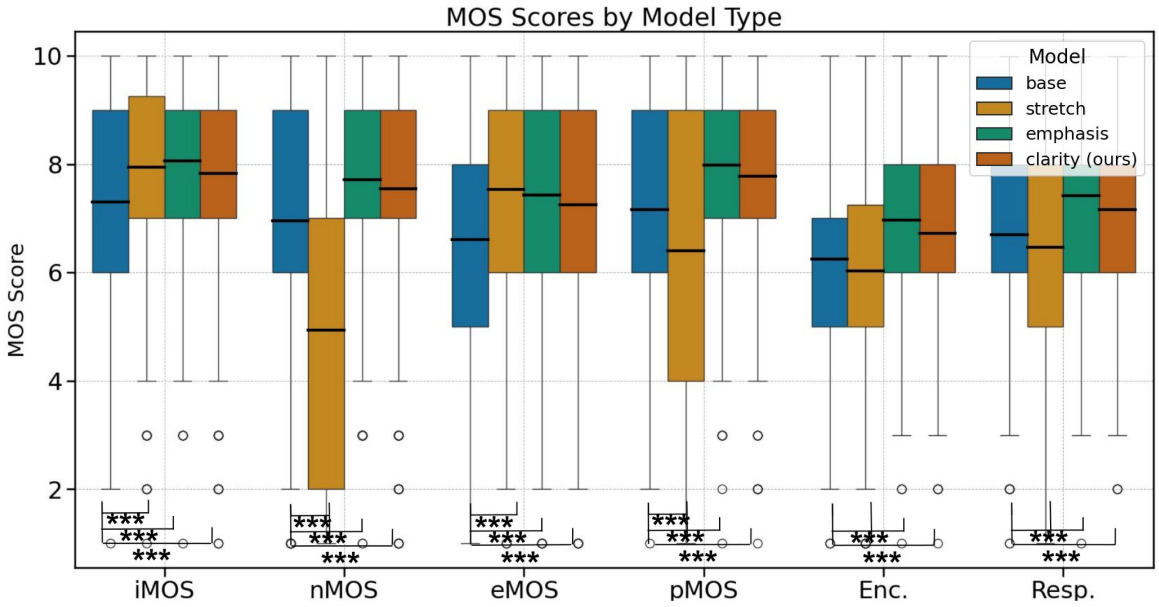

Figure 5.21: Double target word results: Likert MOS scores of words containing tense vowels for intelligibility (iMOS), naturalness (nMOS), effort (eMOS), prosody (pMOS), encouragement (Enc) and respect (Resp.). The baseline is perceived as significantly less intelligible and requiring more listening effort than all of stretch, emphasis and clarity. Baseline and stretch are perceived as significantly less natural with worse prosody than both emphasis and clarity, with stretch lower than baseline. Both baseline and stretch (speaking too fast or too slow) are perceived as significantly less encouraging and respectful.

We used Whisper ASR [145][6] with 72 phrases (generated in the same manner as for the human experiments) and calculated overall WER (WERt) and WER on only the target words. The phrases included those from the human study, and the additional phrases were constructed as in Sec. 5.3.3. We also included the percentage of errors in the target word resulting from minimal tense/lax pair substitution (sub) and what percentage of these substitutions were lax substituted for a tense (t-sub) and tense substituted for a lax (l-sub) vowel. Additionally, we ensured that homophones with the target words were accepted as correct transcriptions.

Similar to L2 speakers, we saw for the "base" TTS, Whisper struggled to predict the correct target words that lack context in the phrase (21.4% vs an expected 5-10% WER for this model), although the performance was slightly higher than that for the L2 speakers (Table 5.8, Fig 5.22). We did not see the same improvements in WER with the "clarity" TTS that we saw in the L2 participants. Instead, we saw an overall slowing down ("stretch") of the TTS decreases the WER in the target words. Yet, we also saw that while the WER on the target words decreased with this slowing down, the proportion of errors in the target word stemming from the minimal pair substitutions was much higher for the "base" TTS

[6] v20231117, medium multilingual

Table 5.8: Whisper ASR results: overall word error rate, target word error rate, tense/lax substitutions, lax substituted for a tense, tense substituted for a lax

TTS Style	**WERt ↓**	**WER ↓**	**sub**	**t-sub**	**l-sub**
Base	17.10%	21.4%	71.42%	61.9%	9.52%
Stretch	**15.98**%	**19.38**%	36.83%	31.57%	5.26%
Emphasis	16.26%	22.4%	31.81%	22.72%	9.09%
Clarity (Ours)	17.68%	24.49%	29.16%	29.16%	0%

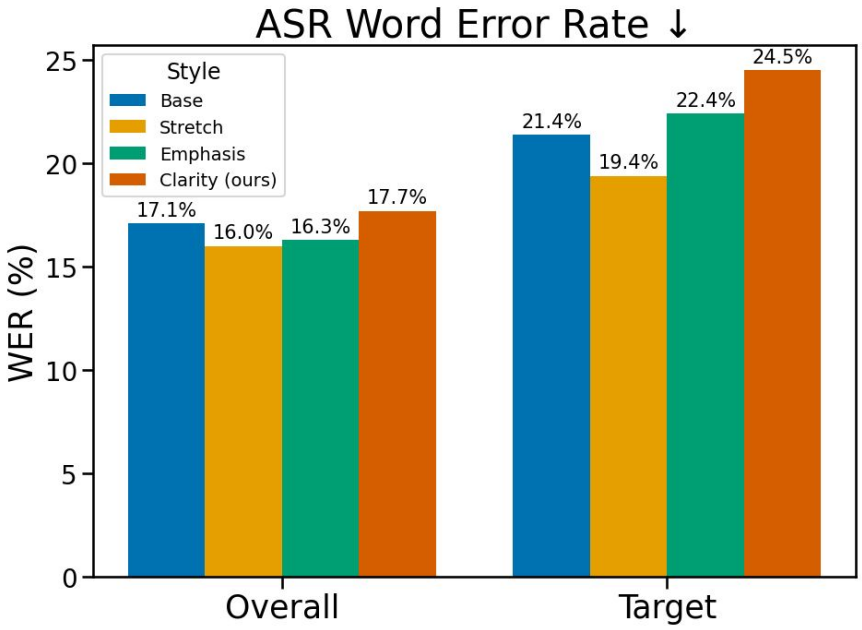

Figure 5.22: Whisper ASR results: We observe that ASR does not align with L2 perception (Fig. 5.20). Stretch had the lowest WER both for the whole phrase (left) and the target words (right).

(71.42%, all other TTS $< 37\%$), while the overall difference in WER both on the whole phrase and the target words was within 3% for all TTS styles. This suggests that while slowing down slightly reduces the overall number of errors, the target words were being predicted as even further from the target, e.g. where "peel" was replaced with "pill" in the "base" TTS was replaced with "peaked" in the "stretch" TTS. Therefore, the ASR does not use the same duration mechanisms as humans when facing difficult predicting words.

5.3.6 Discussion

This study resulted in multiple interesting findings. First, we provided a "clarity mode" as an open-source addition to Matcha-TTS and confirmed that, by applying a stretch to tense vowels for difficult target words, L2 speakers' transcription errors for these words were reduced. Indeed, emphasizing all target words or simply slowing the whole phrase without consideration of the linguistic properties of the vowel reduced the transcription performance of L2 speakers (from French L1 backgrounds), who primarily use duration to determine the

difference between English tense and lax vowels. Moreover, we confirm prior findings that slowing down an entire phrase can be seen as less respectful and encouraging to L2 listeners.

Second, we found that our sample of L2 speakers are unaware they are using this duration mechanism. Through MOS scores, these L2 speakers indicated that they could more easily identify a stretched word, perhaps because they believed they could use the longer vowels (more stable formants) to transcribe a word more easily. This suggests that L2 listeners are limited in their ability to evaluate the effectiveness of adaptive TTS systems, and that objective metrics such as word error rate should serve as a gold standard in future work.

Lastly, we found that Whisper ASR does not use the same duration mechanisms as L2 speakers and, therefore, does not present an adequate replacement for determining the transcription accuracy of synthesized speech for these individuals. This suggests, again, that future development of adaptive TTS systems for L2 speakers must, at the point in time, rely more on data from real human listeners.

Overall, these results have important consequences for TTS assessments, especially for L2 speakers. We cannot simply rely on the same methods used for L1 speaker TTS assessments. Neither user self-rated intelligibility assessments nor automatic systems reflect the true accuracy of difficult vowel perception in L2 speakers, even for those with high proficiency as in our sample, and improved accuracy does not necessarily reflect positive perceptions of the voice (as for "stretch"). Researchers must use both objective and subjective assessments with human participants to ensure they are building inclusive and accessible speech synthesis systems.

Chapter 6

Conclusions and Future Directions

6.1 Previous Recommendations For Robot Voices

In this thesis we made an effort to adhere to the recommendations for spoken language interaction with robots as presented in [180]. There are 25 recommendations organized into seven categories: user experience design; audio processing, speech recognition, and language understanding; speech synthesis and language generation; dialogue; other sensory processes; robustness and adaptability; and infrastructure. We reflect on which relevant recommendations we aimed to achieve in this thesis.

In the category of **user experience design**, we met all but the final recommendation for multi-party interaction. *Focus on language not only as a way to achieve human-like behaviors but also as a way to support limited but highly usable communications abilities.*– We explored and provided recommendations for use cases for our voices both in terms of tasks (Ch. 3 and 5) and physical and social environments (Ch. 4). *Deliberately engineer user perceptions and expectations.*–We focused on making minimal changes to TTS systems and focused on the task at hand, e.g., expressivity (Ch. 3), rather than building natural and human-like voices. This allowed us to ensure we meet our goals for HRI without enforcing unrealistic expectations from the robot. *Work to better characterize the list of communicative competencies most needed for robots in various scenarios.*–Through this work, we strove to understand and define the need for clear and comprehensive voices in specific use cases and for specific users.

We pushed forward all the recommendations for **speech synthesis and language generation**. *Develop the ability to tune speech synthesizers to convey a desired tone, personality, and identity.*–One of the reasons we have chosen Matcha-TTS is because it can be easily controlled for clarity and fine-tuned for expressivity. *Extend the pragmatic repertoire of speech synthesizers.*–We Introduced the use of emojis as an interpretable way to control speech and to represent a robot's internal state (Ch. 3). We also made use of known linguistic properties to improve L2 comprehension (Ch. 5). *Create synthesizers that support real-time control of the voice.*–Once again, Matcha-TTS was chosen for its speed and real-time syn-

thesis abilities. *Develop speech generators that support multimodal interaction.*–In expressive case studies, we have chosen animations that support the communication of expression in our voices (Ch.3).

We addressed all recommendations for **dialogue**. *Focus on highly interactive dialogue.*– We have designed both of our conversational case studies to have short sentences to maintain the flow of the dialogue and user engagement (Ch. 3 and 4). *Make every component able to support real-time responsiveness.*–In our speech-to-speech agent, we made an effort to have each component run in real-time (Ch. 3). However, there is still room for improvement; for further information, see Section 3.4.

Recommendations for other dialogue, sensory processes, robustness and adaptability, audio processing, speech recognition, language understanding, and infrastructure are outside this research's scope and remain open research questions for the community.

6.2 Summary

Our goal was to create a voice that can be used to teach second languages in HRI applications. To achieve this, we first identified a lightweight TTS that could reliably run in real-time, without hallucinations, on a robot platform: Matcha-TTS. We then modified Matcha-TTS to take an emoji prompt to make a temporally expressive TTS to help improve engagement in expressive teaching tasks. We found that this voice is best applied in long-term expressive speech where the variability in the expression is best utilized. Moreover, we provided a toolbox that helps other HRI researchers quickly and easily build their own expressive TTS to their own use case with only 3 minutes of data per expressive style. At the same time, we explored how we could make the voice appropriate and understandable in different social and physical ambient contexts so that the voice can be used in a variety of environments. First, we found that by selecting the correct voice for the environment, we could increase the perception of awareness and appropriateness in the robot. Additionally, for loud environments, a higher pitch could be used as a substitute to increase clarity, where increases in intensity may cause saturation issues in a robot's speakers. Lastly, unlike in human Lombard speech, in a high-energy (e.g., expressive) social ambiance, a large pitch range was preferred alongside this increase in pitch. Finally, we aimed to understand how, from a perception-based perspective, we could improve a TTS for second language comprehension. We explored known linguistic properties of tense and lax vowels in English that are known to be difficult for L2 speakers. We found that, indeed, we could use lengthening of long-tense vowels to aid L2 speakers to avoid confusing these words with their lax minimal pair. In addition, we found that more work needs to be done to make inclusive speech synthesis systems, and in order to assess these systems, we must have humans with a variety of comprehension abilities provide both objective and subjective measures rather than relying on automatic assessments and purely first language speakers.

6.3 Future Work

Future work is required to understand how to combine each of these TTS modifications effectively (see Sec. 6.3.4). In addition, the robustness of each of the methods themselves merit future work.

6.3.1 Expressive voice

For EmojiVoice, we observed, both through the data recording process and the case studies, that although the system seems to be appropriate for long-form interactions with variations from phrase to phrase, we still observe awkwardness in very long phrases. Indeed, humans often will use multiple expressive tones over the course of a long phrase. Our system is limited to one emoji per phrase, and it remains future work to implement, in real-time, multiple tones into a single sentence similar to [95], but in a controllable manner. Moreover, in early explorations, it appeared that this effect was stronger for emojis that may focus on a particular word rather than on an entire phrase, e.g., a "wink" emoji. Future work should explore how emojis can be used as a form of markup for emphasis and tone around specific words in a phrase.

In addition, we found that the high intensity of some of the emotions can cause over-saturation on the robot speakers, which is not noticed when testing the voices through headphones. This was particularly evident in the Miroka trials and may have impacted the results, and is evidence for the importance of testing voices for HRI in person on real robots [181]. The training data was normalized for intensity and new checkpoints are available, eliminating this problem of saturation.

Interestingly, for both case studies, we had 3 participants who preferred the Baseline voice for the interactions and stated that they preferred a flat, robotic-sounding voice for a robot. This shows us that, despite the task, some individuals still want a classic, robotic voice for a robot embodiment. Moreover, all three of these individuals were male, and in general, the male participants seemed to provide lower scores to the expressive voices. Future work should be done to better understand gender preferences for expressive voices and tasks.

6.3.2 Ambient appropriate voice

Although we saw significant improvements in perceptions, most of these were an increase of less than a Likert scale point. Moreover, the average of all of our Likert ratings (social appropriateness, ambiance awareness, comfort, human-likeness, competency) fell between somewhat disagree and agree, suggesting that humans are not perceiving large changes in appropriateness of a voice and are still somewhat uneasy and unopinionated given a robot voice. Further work must be done to expand our abilities to synthesize speech that is adaptable and appropriate if we hope to integrate robots into a multitude of varying everyday spaces.

Real-time systems able to adapt a voice using a sensing-production loop (such as is [29]) are an exciting area for future study; for instance, the robot may assess the distribution of frequencies in the ambiance and adapt its voice to avoid the frequency ranges already filled by background noise. Furthermore, the use of virtual reality could be explored to overcome the limitations in simulating the closed/open characteristics of space, which were not well differentiated in our perception study.

6.3.3 L2 directed TTS

Although we were able to increase the comprehension of difficult words for L2 speakers, our work remains focused on duration as a cue for tense/lax minimal pairs. Further investigations should be launched into how to increase clarity for other vowels, including manipulations to spectral cues, like formants, and later, extending this approach to consonants. For example, manipulating pauses and stress to cue a difficult word may aid L2 speakers to pay attention and hence increase comprehension [182].

Further, through explorations of the results on the participant level in our validation study, we found that the duration mechanism was not clearly universal; rather, it appears to be very strong in certain participants and weak or non-existent in others, which echoes work showing inter-individual differences in cue-weighting [183]. Future work may harness pronunciation data to further customize TTS. Moreover, since our work used forced choice to limit participant responses, a topic of future work is understanding how increasing clarity in a target word affects the rest of the phrase and how robust this duration mechanism is to broader linguistic contexts. Additionally, since our participants had a high level of English proficiency, future work should explore "L2 clarity TTS" with participants of lower proficiency levels, which could explore the modification of words that are less likely to be in the vocabulary of L2 speakers (e.g., "cooed" vs. "could"), using individual vocabularies as a factor predicting perception. Lastly, even more fine-grained control over a TTS, such as in [184], could aid in uncovering other possible mechanisms to aid L2 perception, specifically for how formant control and duration control can be combined.

6.3.4 An ambient-adaptive expressive L2 teaching voice

Each of these elements has been studied individually, and a demo has been created to incorporate all of them simultaneously. In this demo, the robot reads short stories for children using EmojiVoice, and the user can select to hear the voices in L2 clarity mode. Moreover, the pitch increases as the room's noise levels increase. However, an HRI study has not been run to test the interaction of each of these elements in the voice. This remains important future work. For example, when the pitch is increased to maintain clarity in a noisy environment, how does this affect the expressive voice that is already at a higher pitch than a neutral voice? Perhaps this increase needs to be more subtle and more quickly reach a ceiling when using expressive speech. Moreover, we interestingly noticed that the pitch tends

to decrease when the duration lengthening is applied in L2 clarity mode. It is unclear if this is a physical function of the audio manipulation, or rather, if the slower training data also exhibited lower pitch and the model is mimicking slow human voices. As we did not find that pitch played a clear role in manipulating user perception, it remains interesting future work to understand what role, if any, this pitch modification plays in vowel comprehension. Therefore, the next question is how the increase in pitch to reflect environmental clarity interacts with the duration changes for L2 clarity. Lastly, the interaction between the expressive voice and the L2 clarity must be better understood. Each emoji styling has its own unique speech rate, and it may be possible that the L2 clarity duration changes need to change relative to this base speaking rate rather than maintaining a constant multiplicative. For example, if angry speech has a faster speech rate, perhaps a lengthening of 1.6x is not enough to increase clarity in difficult vowels.

Bibliography

[1] E. Ponsot, J. J. Burred, P. Belin, and J.-J. Aucouturier, "Cracking the social code of speech prosody using reverse correlation," *PNAS*, vol. 115, no. 15, pp. 3972–3977, 2018.

[2] W. Mucherah, "Immigrants' perceptions of their native language: Challenges to actual use and maintenance," *Journal of language, identity, and education*, vol. 7, no. 3-4, pp. 188–205, 2008.

[3] M. Guardado, "Loss and maintenance of first language skills: Case studies of hispanic families in vancouver," *The Canadian Modern Language Review*, vol. 58, no. 3, pp. 341–363, 2002.

[4] C. Alphonso, "Pandemic worsens shortage of french-immersion teachers across canada," *The Globe and Mail*. [Online]. Available: https://www.theglobeandmail.com/canada/article-pandemic-worsens-shortage-of-french-immersion-teachers-across-canada

[5] ——, "Shortage prompts school boards to hire teachers who can speak french only slightly better than students, report says." *The Globe and Mail*. [Online]. Available: https://www.theglobeandmail.com/canada/article-school-boards-desperately-short-of-french-language-teachers-report

[6] Truth, R. C. of Canada, and U. Nations, "Truth and reconciliation: Calls to action."

[7] "Blackfoot revitalization project." http://blackfoot-revitalization.cs.sfu.ca, accessed: 2024-12-24.

[8] "The piegan institute." https://www.pieganinstitute.org/our-history, accessed: 2024-12-24.

[9] N. M. Nor and R. A. Rashid, "A review of theoretical perspectives on language learning and acquisition," *Kasetsart Journal of Social Sciences*, vol. 39, no. 1, pp. 161–167, 2018.

[10] N. Randall, "A survey of robot-assisted language learning (rall)," *J. Hum.-Robot Interact.*, vol. 9, no. 1, Dec. 2019.

[11] E. Verhelst, R. Janssens, T. Demeester, and T. Belpaeme, "Adaptive second language tutoring using generative ai and a social robot," in *HRI*. New York, NY, USA: Association for Computing Machinery, 2024, p. 1080–1084.

[12] S. J. Park, J. H. Han, B. H. Kang, and K. C. Shin, "Teaching assistant robot, robosem, in english class and practical issues for its diffusion," in *Advanced Robotics and its Social Impacts*, 2011, pp. 8–11.

[13] G. Gordon, C. Breazeal, and S. Engel, "Can children catch curiosity from a social robot?" in *HRI*. New York, NY, USA: Association for Computing Machinery, 2015, p. 91–98.

[14] R. Nagata, T. Mizumoto, K. Funakoshi, and M. Nakano, "Toward a chanting robot for interactively teaching english to children," in *Second Language Studies: Acquisition, Learning, Education and Technology (L2WS 2010)*, 2010, pp. paper P2–13.

[15] Z.-J. You, C.-Y. Shen, C.-W. Chang, B.-J. Liu, and G.-D. Chen, "A robot as a teaching assistant in an english class," in *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)*, 2006, pp. 87–91.

[16] C.-W. Chang, J.-H. Lee, P.-Y. Chao, C.-Y. Wang, and G.-D. Chen, "Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school," *Journal of Educational Technology and Society*, vol. 13, no. 2, pp. 13–24, 2010.

[17] S. Lee, H. Noh, J. Lee, K. Lee, G. G. Lee, S. Sagong, and M. Kim, "On the effectiveness of robot-assisted language learning," *ReCALL*, vol. 23, no. 1, p. 25–58, 2011.

[18] O. Mubin, S. Shahid, and C. Bartneck, "Robot assisted language learning through games: A comparison of two case studies," *Aust. J. Intell. Inf. Process. Syst.*, vol. 13, 2013.

[19] K. Balkibekov, S. Meiirbekov, N. Tazhigaliyeva, and A. Sandygulova, "Should robots win or lose? robot's losing playing strategy positively affects child learning," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 706–711.

[20] D. E. Takayuki Kanda, Takayuki Hirano and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Human–Computer Interaction*, vol. 19, no. 1-2, pp. 61–84, 2004.

[21] H. Kose and R. Yorganci, "Tale of a robot: Humanoid robot assisted sign language tutoring," in *2011 11th IEEE-RAS International Conference on Humanoid Robots*, 2011, pp. 105–111.

[22] J. K. Westlund, G. Gordon, S. Spaulding, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Learning a second language with a socially assistive robot," *Almere, The Netherlands*, 2015.

[23] J. M. Kory Westlund, L. Dickens, S. Jeong, P. L. Harris, D. DeSteno, and C. L. Breazeal, "Children use non-verbal cues to learn new words from robots as well as people," *International Journal of Child-Computer Interaction*, vol. 13, pp. 1–9, 2017.

[24] E. Martinson and D. Brock, "Improving human-robot interaction through adaptation to the auditory scene," in *HRI*. New York, NY, USA: Association for Computing Machinery, 2007, p. 113–120.

[25] J. Xiao, J. Liu, D. Li, L. Zhao, and Q. Wang, "Speech intelligibility enhancement by non-parallel speech style conversion using cwt and imetricgan based cyclegan," in *MultiMedia Modeling*, B. Þór Jónsson, C. Gurrin, M.-T. Tran, D.-T. Dang-Nguyen, A. M.-C. Hu, B. Huynh Thi Thanh, and B. Huet, Eds. Cham: Springer International Publishing, 2022, pp. 544–556.

[26] S. Novitasari, S. Sakti, and S. Nakamura, "Dynamically adaptive machine speech chain inference for tts in noisy environment: Listen and speak louder," in *Interspeech 2021*, 2021, pp. 4124–4128.

[27] M. Cohn and G. Zellou, "Perception of concatenative vs. neural text-to-speech (tts): Differences in intelligibility in noise and language attitudes," in *Interspeech 2020*, 2020, pp. 1733–1737.

[28] C. Valentini-Botinhao and J. Yamagishi, "Speech enhancement of noisy and reverberant speech for text-to-speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1420–1433, 2018.

[29] Q. Ren, Y. Hou, D. Botteldooren, and T. Belpaeme, "No more mumbles: Enhancing robot intelligibility through speech adaptation," *IEEE Robotics and Automation Letters*, vol. 9, no. 7, pp. 6162–6169, 2024.

[30] D. G. et. al, "Reliability and security of ai hardware," in *ETS*, 2024, pp. 1–10.

[31] L. Kunze, N. Hawes, T. Duckett, M. Hanheide, and T. Krajník, "Artificial intelligence for long-term robot autonomy: A survey," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4023–4030, 2018.

[32] E. Baccour, N. Mhaisen, A. A. Abdellatif, A. Erbad, A. Mohamed, M. Hamdi, and M. Guizani, "Pervasive ai for iot applications: A survey on resource-efficient distributed artificial intelligence," *IEEE Communications Surveys and Tutorials*, vol. 24, no. 4, pp. 2366–2418, 2022.

[33] K. Kinkar, P. Bhosale, A. Kasar, and V. Gutte, "Carbon footprint analysis: Need for green cloud computing," in *ICEARS*, 2022, pp. 1–6.

[34] J. C. Hess, "Chip production's ecological footprint: Mapping climate and environmental impact," in *Interface*, 2024.

[35] L. Manikonda, A. Deotale, and S. Kambhampati, "What's up with privacy? user preferences and privacy concerns in intelligent personal assistants," in *AIES*, 2018, p. 229–235.

[36] M. Abdurachman, E. Abdurachman, F. L. Gaol, and B. Soewito, "Survey on threats and risks in the cloud computing environment," *Procedia Computer Science*, vol. 161, pp. 1325–1332, 2019.

[37] L. Han, H. Guo, Z. Ma, R. Wang, and M. Xiao, "The effect of instructor's voice enthusiasm and visual cueing in multimedia learning," *Journal of Computer Assisted Learning*, vol. 40, no. 6, pp. 3044–3054, 2024.

[38] T. W. Liew, S.-M. Tan, T. M. Tan, and S. N. Kew, "Does speaker's voice enthusiasm affect social cue, cognitive load and transfer in multimedia learning?" *Information and Learning Sciences*, vol. 121, no. 3/4, pp. 117–135, 2020.

[39] W. Liu, "Does teacher immediacy affect students? a systematic review of the association between teacher verbal and non-verbal immediacy and student motivation," *Frontiers in Psychology*, vol. 12, p. 713978, 2021.

[40] J. M. Kory Westlund, S. Jeong, H. W. Park, S. Ronfard, A. Adhikari, P. L. Harris, D. DeSteno, and C. L. Breazeal, "Flat vs. expressive storytelling: Young children's learning and retention of a social robot's narrative," *Frontiers in human neuroscience*, vol. 11, p. 295, 2017.

[41] J. Kennedy, P. Baxter, and T. Belpaeme, "Nonverbal immediacy as a characterisation of social behaviour for human–robot interaction," *International Journal of Social Robotics*, vol. 9, pp. 109–128, 2017.

[42] G. Veletsianos, "The impact and implications of virtual character expressiveness on learning and agent–learner interactions," *Journal of Computer Assisted Learning*, vol. 25, no. 4, pp. 345–357, 2009.

[43] S. Fountoukidou, U. Matzat, J. Ham, and C. Midden, "The effect of an artificial agent's vocal expressiveness on immediacy and learning," *Journal of Computer Assisted Learning*, vol. 38, no. 2, pp. 500–512, 2022.

[44] E. Velner, P. P. Boersma, and M. M. de Graaf, "Intonation in robot speech: Does it work the same as with people?" in *HRI*, 2020, p. 569–578.

[45] S. Hennig and R. Chellali, "Expressive synthetic voices: Considerations for human robot interaction," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 589–595.

[46] O. Niebuhr, J. Voße, and A. Brem, "What makes a charismatic speaker? a computer-based acoustic-prosodic analysis of steve jobs tone of voice," *Computers in Human Behavior*, vol. 64, pp. 366–382, 2016.

[47] L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, E. Gilmartin, C. Murad, C. Munteanu *et al.*, "What makes a good conversation? challenges in designing truly conversational agents," in *CHI*, 2019, pp. 1–12.

[48] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[49] G. Beller, "Expresso: transformation of expressivity in speech," in *Speech Prosody*, 2010, p. paper 043.

[50] P. Tuttösí, S. Mehta, Z. Syvenky, B. Burkanova, G. E. Henter, and A. Lim, "Emojivoice: A tts toolkit for real-time expressive speech on robots." in *Under Review for RA-L*, n.d.

[51] P. Tuttösí, S. Mehta, Z. Syvenky, B. Burkanova, M. Hasafsti, Y. Wang, H. H. Yeung, G. E. Henter, J.-J. Aucouturier, and A. Lim, "Take a look, it's in a book, a reading robot." in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2025.

[52] H. A. C. Maruri, S. Aslan, G. Stemmer, N. Alyuz, and L. Nachman, "Analysis of contextual voice changes in remote meetings," in *Interspeech*, 2021, pp. 2521–2525.

[53] I. Torre, A. B. Latupeirissa, and C. McGinn, "How context shapes the appropriateness of a robot's voice," in *ROMAN*, 2020, pp. 215–222.

[54] S. Ivanov, U. Gretzel, K. Berezina, M. Sigala, and C. Webster, "Progress on robotics in hospitality and tourism: A review of the literature," *J. Hosp. Tour. Technol.*, 2019.

[55] A. Henschel, G. Laban, and E. Cross, "What makes a robot social? a review of social robots from science fiction to a home or hospital near you," *Curr. Robot. Rep.*, vol. 2, 2021.

[56] P. Tuttosi, E. Hughson, A. Matsufuji, C. Zhang, and A. Lim, "Read the room: Adapting a robot's voice to ambient and social contexts," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3998–4005.

[57] E. Hughson, P. Tuttösí, A. Matsufuji, C. Zhang, and A. Lim, "I'm a robot, hear me speak!" in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, p. 909–911.

[58] A. D. Vieira, H. Leite, and A. V. L. Volochtchuk, "The impact of voice assistant home devices on people with disabilities: A longitudinal study," *Technological Forecasting and Social Change*, vol. 184, p. 121961, 2022.

[59] G. A. A. de Oliveira, O. d. F. Oliveira, S. de Abreu, R. W. de Bettio, and A. P. Freire, "Opportunities and accessibility challenges for open-source general-purpose home automation mobile applications for visually disabled users," *Multimedia Tools Appl.*, vol. 81, no. 8, p. 10695–10722, Mar. 2022.

[60] A. J. London, Y. S. Razin, J. Borenstein, M. Eslami, R. Perkins, and P. Robinette, "Ethical issues in near-future socially supportive smart assistants for older adults," *IEEE Transactions on Technology and Society*, vol. 4, no. 4, pp. 291–301, 2023.

[61] W. Seymour, X. Zhan, M. Coté, and J. Such, "A systematic review of ethical concerns with voice assistants," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 131–145.

[62] E. K. McClay, S. Cebioglu, T. Broesch, and H. H. Yeung, "Rethinking the phonetics of baby-talk: Differences across canada and vanuatu in the articulation of mothers' speech to infants," *Developmental science*, vol. 25, no. 2, p. e13180, 2022.

[63] C. Redmon, K. Leung, Y. Wang, B. McMurray, A. Jongman, and J. A. Sereno, "Cross-linguistic perception of clearly spoken english tense and lax vowels based on auditory, visual, and auditory-visual information," *Journal of phonetics*, vol. 81, p. 100980, 2020.

[64] T. Biro, A. J. Olmstead, and N. Viswanathan, "Talker adjustment to perceived communication errors," *Speech communication*, vol. 138, pp. 13–25, 2022.

[65] K. Maniwa, A. Jongman, and T. Wade, "Acoustic characteristics of clearly spoken english fricatives," *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3962–3973, 2009.

[66] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[67] J.-A. Bachorowski, "Vocal expression and perception of emotion," *Current directions in psychological science : a journal of the American Psychological Society*, vol. 8, no. 2, pp. 53–57, 1999.

[68] N. B. Aoki and G. Zellou, "Being clear about clear speech: Intelligibility of hard-of-hearing-directed, non-native-directed, and casual speech for l1- and l2-english listeners," *Journal of Phonetics*, vol. 104, p. 101328, 2024.

[69] K. Rothermich, H. L. Harris, K. Sewell, and S. C. Bobb, "Listener impressions of foreigner-directed speech: A systematic review," *Speech Communication*, vol. 112, pp. 22–29, 2019.

[70] Y. Kita and Y. Kita, "Japanese learners of english and japanese phonology," *Research bulletin of Naruto University of Education*, vol. 34, pp. 209–216, 2019.

[71] P. Inverson, M. Pinet, and B. G. Evans, "Auditory training for experienced and inexperienced second-language learners: Native french speakers learning english vowels," *Applied psycholinguistics*, vol. 33, no. 1, pp. 145–160, 2012.

[72] A. M. Liberman and D. H. Whalen, "On the relation of speech to language," *Trends in cognitive sciences*, vol. 4, no. 5, pp. 187–196, 2000.

[73] C. Stilp, "Acoustic context effects in speech perception," *Wiley interdisciplinary reviews. Cognitive science*, vol. 11, no. 1, pp. e1517–n/a, 2020.

[74] B. McMurray and A. Jongman, "What information is necessary for speech categorization? harnessing variability in the speech signal by integrating cues computed relative to expectations." *Psychological review*, vol. 118, no. 2, p. 219, 2011.

[75] R. S. Newman and J. R. Sawusch, "Perceptual normalization for speaking rate: Effects of temporal distance," *Perception and psychophysics*, vol. 58, no. 4, pp. 540–560, 1996.

[76] E. Reinisch and M. J. Sjerps, "The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context," *Journal of phonetics*, vol. 41, no. 2, pp. 101–116, 2013.

[77] K. Johnson, E. A. Strand, and M. D'Imperio, "Auditory–visual integration of talker gender in vowel perception," *Journal of phonetics*, vol. 27, no. 4, pp. 359–384, 1999.

[78] P. Tuttösí, H. H. Yeung, Y. Wang, F. Wang, G. Denis, J.-J. Aucouturier, and A. Lim, "Mmm whatcha say? uncovering distal and proximal context effects in first and second-language word perception using psychophysical reverse correlation," in *Interspeech*, 2024, pp. 1010–1014.

[79] P. Tuttösí, H. H. Yeung, Y. Wang, F. Wang, J.-J. Aucouturier, and A. Lim, "You're sounding a little tense: L2 tailored tts using durational vowel properties," in *Under Review for Interspeech 2025*, n.d.

[80] Coqui-AI, "Coquixtts," *https://github.com/coqui-ai/TTS/tree/dev*, 2021.

[81] P. Peng, P.-Y. Huang, A. Mohamed, and D. Harwath, "Voicecraft: Zero-shot speech editing and text-to-speech in the wild," *arXiv*, 2024.

[82] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," *arXiv*, 2024.

[83] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP*, 2021, pp. 6588–6592.

[84] Y. Ju, I. Kim, H. Yang, J.-H. Kim, B. Kim, S. Maiti, and S. Watanabe, "Trinitts: Pitch-controllable end-to-end tts without external aligner," in *Interspeech*, 2022, pp. 16–20.

[85] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv*, 2022.

[86] D. Diatlova and V. Shutov, "Emospeech: guiding fastspeech2 towards emotional text to speech," in *SSW*, 2023, pp. 106–112.

[87] E. Strickland, D. Aubakirova, D. Doncenco, D. Torres, and M. Evrard, "Naijatts: A pitch-controllable tts model for nigerian pidgin," in *SSW*, 2023, pp. 248–249.

[88] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-TTS: A fast TTS architecture with conditional flow matching," in *ICASSP*, 2024.

[89] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023.

[90] D.-H. Cho, H.-S. Oh, S.-B. Kim, S.-H. Lee, and S.-W. Lee, "Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech," in *Interspeech 2024*, 2024, pp. 1810–1814.

[91] T. Bott, F. Lux, and N. T. Vu, "Controlling emotion in text-to-speech with natural language prompts," in *Interspeech*, 2024, pp. 1795–1799.

[92] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for storytelling applications," *Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1137–1144, 2006.

[93] K. Gustafson and D. House, "Fun or boring? a web-based evaluation of expressive synthesis for children," in *Eurosepeech*, 2001, pp. 565–568.

[94] Y. Xiao, S. Zhang, X. Wang, X. Tan, L. He, S. Zhao, F. K. Soong, and T. Lee, "Contextspeech: Expressive and efficient text-to-speech for paragraph reading," in *Interspeech 2023*, 2023, pp. 4883–4887.

[95] S. Liu, Y. Guo, X. Chen, and K. Yu, "Storytts: A highly expressive text-to-speech dataset with rich textual expressiveness annotations," in *ICASSP*, 2024, pp. 11 521–11 525.

[96] A. R. Bradlow, "Confluent talker-and listener-oriented forces in clear speech production," *Laboratory phonology*, vol. 7, pp. 241–273, 2002.

[97] D. Pelegrin-Garcia, B. Smits, J. Brunskog, and C.-H. Jeong, "Vocal effort with changing talker-to-listener distance in different acoustic environments." *J. Acoust. Soc*, vol. 129 4, pp. 1981–90, 2011.

[98] V. Hazan and R. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," *J. Acoust. Soc*, vol. 130, pp. 2139–52, 2011.

[99] J. A. Caballero, N. Vergis, X. Jiang, and M. D. Pell, "The sound of im/politeness," *Speech Commun.*, vol. 102, pp. 39–53, 2018.

[100] D. Burnham, C. Kitamura, and U. Vollmer-Conna, "What's new, pussycat? on talking to babies and animals," *Science*, vol. 296, p. 1435, 2002.

[101] E. A. Piazza, M. C. Iordan, and C. Lew-Williams, "Mothers consistently alter their unique vocal fingerprints when communicating with infants," *Current Biology*, vol. 27, no. 20, pp. 3162–3167, 2017.

[102] B. De Boer and P. K. Kuhl, "Investigating the role of infant-directed speech with a computer model," *Acoustics Research Letters Online*, vol. 4, no. 4, pp. 129–134, 2003.

[103] K. Hirsh-Pasek and R. Treiman, "Doggerel: Motherese in a new context," *Journal of child language*, vol. 9, no. 1, pp. 229–237, 1982.

[104] C. Lam and C. Kitamura, "Mommy, speak clearly: induced hearing loss shapes vowel hyperarticulation." *Dev. Sci.*, vol. 15, no. 2, pp. 212–21, 2012.

[105] C. Mayo, V. Aubanel, and M. Cooke, "Effect of prosodic changes on speech intelligibility," in *Interspeech*, vol. 2, 2012.

[106] S. D. Craig and N. L. Schroeder, "Text-to-speech software and learning: Investigating the relevancy of the voice effect," *J. Educ. Comput. Res.*, vol. 57, no. 6, pp. 1534–1548, 2019.

[107] R. Vipperla, S. Park, K. Choo, S. Ishtiaq, K. Min, S. Bhattacharya, A. Mehrotra, A. G. C. P. Ramos, and N. D. Lane, "Bunched lpcnet : Vocoder for low-cost neural text-to-speech systems," 2020, [Online] Available:arXiv:2008.04574.

[108] K.-G. Oh, C.-Y. Jung, Y.-G. Lee, and S.-J. Kim, "Real-time lip synchronization between text-to-speech (tts) system and robot mouth," in *ROMAN*, 2010, pp. 620–625.

[109] A. Matsufuji and A. Lim, "Perceptual effects of ambient sound on an artificial agent's rate of speech," in *Companion of HRI*, 2021, pp. 67–70.

[110] Y. Okuno, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Providing route directions: Design of robot's utterance, gesture, and timing," in *HRI*, 2009, pp. 53–60.

[111] A. Hönemann and P. Wagner, "Adaptive speech synthesis in a cognitive robotic service apartment: An overview and first steps towards voice selection," in *ESSV*, 2015.

[112] S. J. Sutton, P. Foulkes, D. Kirk, and S. Lawson, "Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction," in *CHI*, 2019, p. 1–14.

[113] N. Lubold, E. Walker, and H. Pon-Barry, "Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion," in *HRI*, 2016, pp. 255–262.

[114] K. Fischer, L. Naik, R. M. Langedijk, T. Baumann, M. Jelínek, and O. Palinko, "Initiating human-robot interactions using incremental speech adaptation," in *Companion of HRI*, 2021.

[115] A. Hayamizu, M. Imai, K. Nakamura, and K. Nakadai, "Volume adaptation and visualization by modeling the volume level in noisy environments for telepresence system," in *HAI*. ACM, 2014.

[116] P. K. Kuhl, J. E. Andruski, I. A. Chistovich, L. A. Chistovich, E. V. Kozhevnikova, V. L. Ryskina, E. I. Stolyarova, U. Sundberg, and F. Lacerda, "Cross-language analysis of phonetic units in language addressed to infants," *Science*, vol. 277, no. 5326, pp. 684–686, 1997.

[117] A. J. Watkins and S. J. Makin, "Perceptual compensation for speaker differences and for spectral-envelope distortion," *The Journal of the Acoustical Society of America*, vol. 96, no. 3, pp. 1263–1282, 1994.

[118] J. Miller and A. Liberman, "Some effects of later-occurring information on the perception of stop consonant and semivowel," *Perception and Psychophysics*, vol. 25, no. 6, pp. 457–465, 2023.

[119] J. Chen, T. Baer, and B. C. Moore, "Effect of enhancement of spectral changes on speech intelligibility and clarity preferences for the hearing impaired," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2987–2998, 2012.

[120] S. Kang, K. Johnson, and G. Finley, "Effects of native language on compensation for coarticulation," *Speech Communication*, vol. 77, pp. 84–100, 2016.

[121] V. A. Mann, "Distinguishing universal and language-dependent levels of speech perception: Evidence from japanese listeners' perception of english "l" and "r"," *Cognition*, vol. 24, no. 3, pp. 169–196, 1986.

[122] N. Viswanathan, J. S. Magnuson, and C. A. Fowler, "Similar response patterns do not imply identical origins: an energetic masking account of nonspeech effects in compensation for coarticulation." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 4, p. 1181, 2013.

[123] C. C. Heffner, L. C. Dilley, D. J. McAuley, and M. A. Pitt, "When cues combine: How distal and proximal acoustic cues are integrated in word segmentation," *Language and Cognitive Processes*, vol. 28, no. 9, pp. 1275–1302, 2013.

[124] K. B. Shatzman and J. M. McQueen, "Prosodic knowledge affects the recognition of newly acquired words," *Psychological Science*, vol. 17, no. 5, pp. 372–377, 2006.

[125] A. A. Jr and J. Lovell, "Stimulus features in signal detection," *The Journal of the Acoustical Society of America*, vol. 49, no. 6B, pp. 1751–1756, 1971.

[126] S. Yan, C. Soladié, J.-J. Aucouturier, and R. Seguier, "Combining gan with reverse correlation to construct personalized facial expressions," *PloS one*, vol. 18, no. 8, pp. e0 290 612–e0 290 612, 2023.

[127] L. Goupil, E. Ponsot, D. Richardson, G. Reyes, and J.-J. Aucouturier, "Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature," *Nature communications*, vol. 12, no. 1, pp. 861–861, 2021.

[128] R. F. Murray, "Classification images: A review," *Journal of vision*, vol. 11, no. 5, pp. 2–2, 2011.

[129] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019, sound.

[130] V. P. et. al, "Scaling up online speech recognition using convnets," in *Interspeech 2020*, 2020, pp. 3376–3380.

[131] Q. Bai, Q. Dan, Z. Mu, and M. Yang, "A systematic review of emoji: Current research and future perspectives," *Frontiers in Psychology*, vol. 10, 2019.

[132] M. Schröder, H. Pirker, and M. Lamolle, "First suggestions for an emotion annotation and representation language," in *LREC*, vol. 6, 2006, pp. 88–92.

[133] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor, "Emotion knowledge: Further exploration of a prototype approach," *Journal of personality and social psychology*, vol. 52, no. 6, pp. 1061–1086, 1987.

[134] R. Plutchik, "Chapter 1 - a general psychoevolutionary theory of emotion," in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds. Academic Press, 1980, pp. 3–33.

[135] T. Dimson, "Emojineering part 1: Machine learning for emoji trends," *Instagram Engineering Blog*, vol. 30, 2015.

[136] G. Savage and A. Yung, "Real-time message sentiment augmentation by emoji symbols," 2024.

[137] M. Dunn and K. Hopkinson, "How good is gpt's "emojinal intelligence"? investigating emoji patterns in llm-generated social media text," in *Proceedings of the International Conference on AI Research*. Academic Conferences and publishing limited, 2024.

[138] L. Grassi, Z. Hong, C. T. Recchiuto, and A. Sgorbissa, "Grounding conversational robots on vision through dense captioning and large language models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 5492–5498.

[139] R. Janssens, P. Wolfert, T. Demeester, and T. Belpaeme, "Integrating visual context into language models for situated social conversation starters," *IEEE Transactions on Affective Computing*, pp. 1–14, 2024.

[140] C. Zhang, J. Chen, J. Li, Y. Peng, and Z. Mao, "Large language models for human–robot interaction: A review," *Biomimetic Intelligence and Robotics*, vol. 3, no. 4, p. 100131, 2023.

[141] E. Billing, J. Rosén, and M. Lamb, "Language models for human-robot interaction," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 905–906.

[142] C. B. et. al, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[143] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, 2018.

[144] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2019.

[145] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[146] M. Bilac, M. Chamoux, and A. Lim, "Gaze and filled pause detection for smooth human-robot conversations," in *Humanoids*, 2017, pp. 297–304.

[147] A. D. et al., "The llama 3 herd of models," 2024.

[148] J. R. Lewis and IBM.Corp, "Investigating mos-x ratings of synthetic and human voices," *Voice Interaction Design*, vol. 2, no. 1, p. 22, 2018.

[149] Y. Korotkova, I. Kalinovskiy, and T. Vakhrusheva, "Word-level text markup for prosody control in speech synthesis," in *Interspeech*, 2024, pp. 2280–2284.

[150] P. van Rijn, S. Mertes, K. Janowski, K. Weitz, N. Jacoby, and E. André, "Giving robots a voice: Human-in-the-loop voice creation and open-ended labeling," in *CHI*, 2024.

[151] T. Raitio, J. Latorre, A. Davis, T. Morrill, and L. Golipour, "Improving the quality of neural tts using long-form content and multi-speaker multi-style modeling," in *SSW*, 2023.

[152] P. Arias, P. Belin, and J.-J. Aucouturier, "Auditory smiles trigger unconscious facial imitation," *Current Biology*, vol. 28, no. 14, pp. R782–R783, 2018.

[153] E. Lombard, "Le signe de l'élévation de la voix," *Ana. d. Mal. de L'Oreillexdu du larynx [etc]*, vol. 37, pp. 101–119, 1911.

[154] A. Castellanos, J.-M. Benedí, and F. Casacuberta, "An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect," *Speech Commun.*, vol. 20, no. 1, pp. 23–35, 1996.

[155] M. Mller, *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer Publishing Company, Inc., 2015.

[156] H. F. Wertzner, S. Schreiber, and L. Amaro, "Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders," *Braz j. of otorh*, vol. 71, no. 5, pp. 582–588, 2005.

[157] T. IR and P. A, "Vocal loudness variation with spectral slope," *J. Speech Lang. Hear*, vol. 63, no. 1, pp. 74–82, 2020.

[158] A. Lerch, *Instantaneous Features*. John Wiley and Sons, Ltd, 2012, ch. 3, pp. 31–69.

[159] D. Kewley-Port, O.-S. Bohn, and K. Nishi, "The influence of different native language systems on vowel discrimination and identification," *The Journal of the Acoustical Society of America*, vol. 117, no. 4_Supplement, pp. 2399–2399, 2005.

[160] F. Han, "Pronunciation problems of chinese learners of english." *Ortesol Journal*, vol. 30, pp. 26–30, 2013.

[161] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[162] M. Monnot, "La prononciation du français contemporain," *The French Review*, vol. 48, no. 1, pp. 284–285, 1974.

[163] W. Strange, A. Weber, E. S. Levy, V. Shafiro, M. Hisagi, and K. Nishi, "Acoustic variability within and across german, french, and american english vowels: Phonetic context effects," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1111–1129, 2007.

[164] J. L. Miller, M. Mondini, F. Grosjean, and J.-Y. Dommergues, "Dialect effects in speech perception: The role of vowel duration in parisian french and swiss french," *Language and speech*, vol. 54, no. 4, pp. 467–485, 2011.

[165] S. F. ul Hassan, "The nature of stress in english language: a study from a perspective of rule-governed approach," *Language in India*, vol. 12, no. 12, p. 362, 2012.

[166] D. Frost, "The perception of word stress in english and french: Which cues for native english and french speakers?" in *English Pronunciation: Issues and Practices*, 2009, pp. 57–73.

[167] Y.-A. Lu and S.-I. Lee-Kim, "The effect of linguistic experience on perceived vowel duration: Evidence from taiwan mandarin speakers," *Journal of Phonetics*, vol. 86, p. 101049, 2021.

[168] Y. Hirata, "Effects of speaking rate on the vowel length distinction in japanese," *Journal of Phonetics*, vol. 32, no. 4, pp. 565–589, 2004.

[169] T. Kozasa, "The interaction of duration and pitch in japanese long vowels," in *Annual Meeting of the Berkeley Linguistics Society*, 2004, pp. 211–222.

[170] J. E. Flege, "The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification," *Journal of phonetics*, vol. 15, no. 1, pp. 47–65, 1987.

[171] E. S. Levy, "On the assimilation-discrimination relationship in american english adults' french vowel learning," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2670–2682, 2009.

[172] J. L. Sturm, "Explicit phonetics instruction in l2 french: A global analysis of improvement," *System*, vol. 41, no. 3, pp. 654–662, 2013.

[173] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," version 6.4.06, retrieved 25 February 2024 from http://www.praat.org/.

[174] J. J. Burred, E. Ponsot, L. Goupil, M. Liuni, and J.-J. Aucouturier, "Cleese: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition," *PloS one*, vol. 14, no. 4, p. e0205943, 2019.

[175] R. Adolphs, L. Nummenmaa, A. Todorov, and J. V. Haxby, "Data-driven approaches in the investigation of social perception," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1693, p. 20150367, 2016.

[176] A. Osses, E. Spinelli, F. Meunier, E. Gaudrain, and L. Varnet, "Prosodic cues to word boundaries in a segmentation task assessed using reverse correlation," *JASA Express Letters*, vol. 3, no. 9, 2023.

[177] A. Joly, M. Nicolis, E. Peterova, A. Lombardi, A. Abbas, A. van Korlaar, A. Hussain, P. Sharma, A. Moinet, M. Lajszczak, P. Karanasou, A. Bonafonte, T. Drugman, and E. Sokolova, "Controllable emphasis with zero data for text-to-speech," in *12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 113–119.

[178] G. Ma, P. Hu, J. Kang, S. Huang, and H. Huang, "Leveraging phone mask training for phonetic-reduction-robust e2e uyghur speech recognition," in *Interspeech 2021*, 2021, pp. 306–310.

[179] J. Taylor and K. Richmond, "Confidence intervals for asr-based tts evaluation," in *Interspeech 2021*, 2021, pp. 2791–2795.

[180] M. M. et. al, "Spoken language interaction with robots: Recommendations for future research," *Computer Speech and Language*, vol. 71, p. 101255, 2022.

[181] S. Wang, Éva Székely, and J. Gustafson, "Contextual interactive evaluation of tts models in dialogue systems," in *Interspeech*, 2024, pp. 2965–2969.

[182] N. Nagata, "The effects of silent pauses on listening comprehension: a case of japanese learners of english as a foreign language," *Doctoral dissertation, Waseda University*, 2002.

[183] J. Schertz, T. Cho, A. Lotto, and N. Warner, "Individual differences in phonetic cue use in production and perception of a non-native sound contrast," *Journal of phonetics*, vol. 52, pp. 183–204, 2015.

[184] C. Tånnander, S. Mehta, J. Beskow, and J. Edlund, "Beyond graphemes and phonemes: continuous phonological features in neural text-to-speech synthesis," in *Interspeech 2024*, 2024, pp. 2815–2819.

Appendix A

Ambiance Zoom Experiment Script

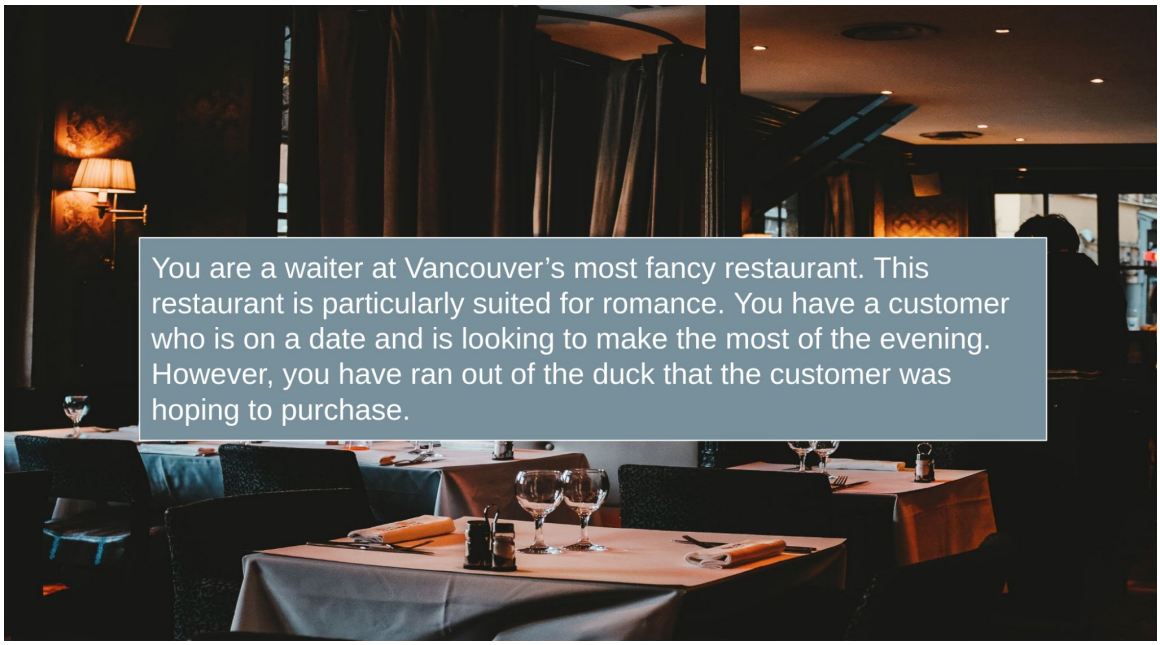

Figure A.1: Zoom background image and prompt for the "fancy restaurant" ambiance in the Chapter 3.

Waiter: Hi there, I hope you're doing well. I hope you have had a chance to look at the menu. I thought I would let you know, today we do have steak frites as our special. Anyways what can I start you off with?

Customer: Can I get a bottle of your finest Prosecco?

Waiter: Sure! I can definitely do that!

Customer: Wait! Can you also make sure the Prosecco is nice and cold and served in champagne glasses?

Waiter: Of course, we can definitely do that. Would you also like some water for the table?

Customer: Hmm. What kind of water do you have?

Waiter: We have still, sparkling, and tap water.

Customer: We will go with the sparking, please!

Waiter: Alrighty, then! I'll be right back with your Prosecco with two champagne glasses and sparkling water!

A few minutes pass by...

Waiter comes and brings customer order... Waiter: Hi again, here is your prosecco. Are you both ready to order food?

Customer: Thanks. I think we would both like to have the Duck?

Waiter: Oh no, unfortunately we just ran out of the Duck! Is there anything else that you see?

Customer: How do you not have any Duck? That's your staple item?

Waiter: Uh, it has been a pretty hectic evening; busier than normal!

Customer: Hmm. Fine, I guess we'll just have the steak frites then.

A few minutes pass by...

Waiter returns with order. Waiter: Here are your steak frites. I hope you enjoy. Let me know if there is anything else I can get you.

Customer: I am not going to lie to you, I would have much rather preferred the Duck. But the frites is good enough.

Waiter: I am so sorry we couldn't get you the Duck. I am sure next time we will have more available.

Customer: Ya, it is a bit odd. Usually I thought businesses like this are more prepared and make enough items just in case.

Anyways, can we get the bill?

Waiter: For sure, I will be right back with your bill!

Appendix B

Battery for Speaker Ambiance Validation study

- This person's voice is socially appropriate for the scene.
- This person knows how they should present themselves in this ambient context.
- This person is aware of the surrounding ambiance.
- This voice sounds like it was originally recorded in this ambiance.
- This person knows how to adapt their voice this ambiance.
- This person makes me feel comfortable.
- I feel uneasy listening to this person.
- I feel at easy when conversing with this person.
- It took a great deal of effort to understand what this person was saying.
- Certain words were difficult to understand.
- This person's speech was very clear.

Appendix C

L2 TTS Experiment Word List

The word cut seemed important to the instructions. She kept mentioning cot during the conversation. The speaker mentioned pill, or at least something similar. The word pill was what she was trying to write. The phrase had fool somewhere in the middle of it. I saw full written on the note pad. The sign mentioned sin, but the person said scene. He wrote down bought, but remembered it as but. In his talk he kept using could, but I am pretty sure he meant cooed. The paper mentioned kid, yet he is telling me knot. There was confusion between pull and bean in their speech. I am not sure if the word was pool or if cup was the right one. Sheep goes on the top of the page and dull goes on the bottom. Bit was the first word he said, then nut followed. Actually hut is the correct word, it was replaced with should by accident. Maybe he said hot, but I really thought keyed was what he said. Reap was a more important word in the story than wooed.

Appendix D

L2 TTS Experiment Word List - Whisper ASR

Ship was what I heard, but maybe I misunderstood.
Sheep might have been what they meant, but it wasn't clear.
Pool was the word I caught, though it could've been something else.
Pull might have been what they were talking about, but I'm unsure
Cut was the word used, but it seemed odd in context.
Cot could have been the intention, but I wasn't sure.
I thought they said something about peel.
I think the phrase ended with full.
The word on the sign was pill.
No, the text message definitely said fool.
I thought they ended their sentence with rut.
It sounded like the last word they said was rot.
The word bean seemed important to the instructions.
She kept mentioning bin during the conversation.
The speaker mentioned wood, or at least something similar.
The word wooed was what she was trying to write.
The phrase had dull somewhere in the middle of it.
I saw doll written on the note pad.
Doll was what I heard, but maybe I misunderstood.
Dull might have been what they meant, but it wasn't clear.
Wood was the word I caught, though it could've been something else.
Wooed might've been what they were talking about, but I'm unsure.
Bean was the word used, but it seemed odd in context.
Bin could've been the intention, but I wasn't sure.
I thought they said something about rot.
I think the phrase ended with rut.
The word on the sign was ship.
No, the text message definitely said sheep.
I thought they ended their sentence with pool.
It sounded like the last word they said was pull.

The word cut seemed important to the instructions.
She kept mentioning cot during the conversation.
The speaker mentioned peel, or at least something similar.
The word pill was what she was trying to write.
The phrase had fool somewhere in the middle of it.
I saw full written on the note pad.
The sign mentioned sin, but the person said scene.
The sign mentioned scene, but the person said sin.
There was confusion between cup and cop in their speech.
There was confusion between cop and cup in their speech.
He wrote down luke, but remembered it as look.
He wrote down look, but remembered it as luke.
The paper mentioned bit, yet he is telling me knot.
The paper mentioned beat, yet he is telling me nut.
I am not sure if the word was could or if keyed was the right one.
I am not sure if the word was cooed or if kid was the right one.
Now note down heat and the word hut on the bottom.
Now note down hit and the word hot and the bottom.
The sign mentioned could, but the person said cooed.
The sign mentioned cooed, but the person said could.
There was confusion between kid and keyed in their speech.
There was confusion between keyed and kid in their speech.
He wrote down hut, but remembered it as hot.
He wrote down hot, but remembered it as hut.
The paper mentioned nut, yet he is telling me heat.
The paper mentioned knot, yet he is telling me hit.
I am not sure if the word was scene or if cup was the right one.
I am not sure if the word was sin or if cop was the right one.
Now note down beat and the word look on the bottom.
Now note down bit and the word luke on the bottom.
He whispered about hearing someone say pill and cup nearby.
He whispered about hearing someone say peel and cop nearby.
The presenter said dull but I think he meant but.
The presenter said doll but I think he meant bought.
Please write down the words reap and rot on your paper.
Please write down the words rip and rut on your paper.
He whispered about hearing someone say but and rip nearby.
He whispered about hearing someone say bought and reap nearby.
The presenter said peel but I think he meant rot.
The presenter said pill but I think he meant rut.
Please write down the words cup and dull on your paper.
Please write down the words cop and doll on your paper.