Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Engineering

Department of Electronic and Electrical Engineering

Segmental evaluation of Text-to-Speech Synthesizers

Ayushi Pandey

April 24, 2024

A dissertation submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Declaration

I hereby declare that this thesis is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: ______________________ Date: ______________________

Abstract

Advancements in speech synthesis technology have mandated the need for reliable methods for its evaluation. Present day evaluation, dominated by subjective listening tests, provides at best, a general overall picture of the perceived speech quality. It does not provide information about the relationship between acoustic parameters, and their contribution to perceived attributes of synthetic speech such as naturalness, similarity and pleasantness. Naturalness in particular, which is a widely used standard in synthetic speech evaluation, is often underspecified. It has also been reported that factors like modified instructions, contextual framing, or user-expectations with the application of synthetic speech can influence the ratings of naturalness. However, we see evidence of consistent listener agreement on their ratings of naturalness, in multiple studies of synthetic speech evaluation. This leads us to hypothesize, that there may be information in the acoustic signature of Text-To-Speech (TTS) signals that the listeners exploit to make a judgment on naturalness.

The primary goal of this thesis is to use contrastive properties of speech segments present in corpora of synthetic speech for evaluating the naturalness of synthetic speech. The concept of naturalness has been discussed as a multi-faceted perceptual attribute. The scope of this thesis is limited to one aspect: the human-likeness of TTS voices. We have selected the Blizzard Challenge 2013 (BC-2013) corpus for our analysis, because it provides parallel TTS data over a wide selection of Hidden Markov Model (HMM), Unit-selection, Hybrid and more recently neural TTS techniques. Contrastive features of vowels and obstruent consonants are extracted using standard acoustic-phonetic and corpus phonetics techniques. Features of each synthetic voice are compared with the human voice, which is held as the reference. Then, a new subjective evaluation framework is proposed which complements the diagnostic nature of the segmental analysis.

Our results show that segmental evaluation can be used to provide diagnostic information that is often missed by traditional subjective tests. In non-neural systems, we find features of obstruent consonants such spectral tilt and RMS amplitude can be useful for identifying quality-differences between systems and groups of systems. Additionally, vowel features such as within-category dispersion showed an above chance correlation ($\rho = 0.64$) with the perceived Mean Opinion Score (MOS). Next, we show that segmental evaluation can be extended successfully to evaluating modern, neural TTS synthesizers. First, we find that neural TTS performs very well in modelling vowels, and has improved over several features of the older, non-neural TTS synthesizers. Only a few features like F0 onset and spectral tilt show statistically significant deviations from the human voice. However, features of voiceless obstruents were found to be distorted, i.e, they deviated significantly from the reference human voice. This is one of the major findings of this thesis. We also investigate the perceptual significance of the deviation in obstruents, through a novel subjective evaluation design. The study involved presenting stimuli of varying lengths to 128 participants, who were asked to identify whether each stimulus was produced by a human or a machine. We hypothesized that the length of the stimuli would aid in more accurate detection between human and machine stimuli. The participants' responses were captured using a 2-alternative forced choice task, and were analyzed using a logistic regression. In obstruent-rich stimuli, we indeed found a 22.37% increase in accuracy as length increased with strongly significant effects (p-val < 0.001).

The findings in this thesis can be used to provide localized insights into feature distortion, and can be extended to provide real-time feedback for TTS engineers. These findings also highlight the usefulness of phonetics in TTS technology, and enable greater interaction between the communities.

Acknowledgements

My first gratitude is to my supervisor Professor Naomi Harte, a role model in every imaginable aspect of life. She created an equal, inclusive space for a linguist in an AI lab, and she has fought for the work that originates in my own identity. Every contribution in this PhD is a result of this rare courage, that few scientists of her stature can maintain. Her faith in me has had a transformative power: I am now a little less afraid. Thank you, Naomi. Next, no part of this research was possible without my co-supervisor Dr Sébastien Le Maguer. His rock-solid encouragement made me understand that the hard phases of research are just in fact, procedural. Thank you Séb, for keeping me safe from my own head. I also thank Professor Jens Edlund for his sustained faith, which reached me even through a thousand miles apart. Thank you Jens, for getting me out of some very tight spots. I also thank Professor Julie Carson-Berndsen for extremely valuable advice in the initial stages of this PhD. Finding curious and brave advisors is the most profound discovery of this PhD.

This thesis stands on the shoulders of several phoneticians and computer scientists, with whom I was fortunate to interact. The most important gratitude is extended to my first advisor Professor Indranil Dutta who visualized the path of phonetics research for me, and was instrumental in helping me launch. Then, I thank Professor Anish Koshy, whose lectures inspired me to pursue the fascinating, all-consuming field of linguistics. I thank the faculty at IIIT Hyderabad, to whom I owe my computational skills. I thank Charlie Redmon for his indispensable advice on obstruents. I thank Eleanor Chodroff and Professor Kevin Tang, for contributing their expertise on corpus phonetics. Finally, I thank Professor Yvette Graham and Dr Rob Clark for their insightful and detailed feedback as my PhD examiners.

I take pride in being part of supportive workplace, which calms and strengthens the mind. First of all, I thank Professor Anil Kokaram for fostering bonds within SigMedia that always felt familial. I extend my gratitude to my colleagues and friends - George, Matt, Mark, Ed, Sam, Sam, Xin, Vibhoothi, Clément, Megan, Giulia, Darren, Francois and DJ - for the favourite hours of lunch. Your presence meant that things were going to be alright for a while. I wish Zhaofeng, Udit, Boldo, JZ and Conall who are at the beginning of this path, the very best of luck. I thank Harm and Ambika, my KTH friends, for making international conferences feel a little less foreign. And I thank all the members of the lab - Mark, Shane,

John, Michael, Caroline, Eugene and specially to Stephen Carroll - who go out of their way to ensure smooth functioning within the lab. I thank The Campanile whose sight empowered me every morning.

It is a privilege to have lived in Ireland, whose majestic beauty and the kindness of its people baffles me even today. Thank you Maria Hayden for creating a space for me in your smiling, sunlit home. And thank you for sharing a hundred cakes, chocolates and tidbits of Irish life that I would have never known otherwise. I'm also grateful to have found Shishir Bhaiya and Manu Bhabhi, my best friends and companions in Dublin. Their presence solves the hardest problems for me. I thank my Bhanji Vedu for brightening my life like never before and running the cool end of Baby and Bbbboooo! I also thank Neil and Quinton, for organizing Improv meetups at The Playground. I laugh so loud there, and I'm sure I delayed my thesis after I found you.

The most important acknowledgment extends to my family, whose love and sacrifice have written the best parts of this thesis. Thank you Ma for being my Pratham Guru: for introducing me to excellence, and for making do with pride when Diwali came. Thank you Papa for your infinite faith in the decisions of my life: your trust is both my feet and my path. Thanks to my Didi-Motz: your image in my head grinds every scary monster into powder. Thank you Mamu, MamiJi, Mausi, MausaJi, Chachi and ChachaJi, my brother Munmun and 7 sisters (Avu Didi, Amu Didi, Somya Didi, Kanchu Didi, Bitz, Shubhu, Runjhun) who are made of pure, sparkling starlight. A special thanks to my parents-in-law MummyJi and PapaJi who bear a ton of social pressures on my behalf, and never let me find out. Finally, I thank my husband Brij Mohan who was the warmest hand for me to hold without conditions. I only got to see him once this year. This thesis is dedicated to the lives and families of non-European PhD researchers.

Contents

List of Figures

List of Tables

List of acronyms

ACR Absolute Category Rating

BC-2013 Blizzard Challenge 2013

CBHG 1-D convolution bank + highway network + bidirectional gated recurrent unit

CLID Cluster Identification Test

CNC Consonant Nucleus Consonant

DCR Degradation Category Ratings

DNN deep neural network

DRT Diagnostic Rhyme Test

EEG ElectroEncephaloGram

FFTr Feed Forward Transformer

GAN Generative Adversarial Assistants

HMM Hidden Markov Model

IAF Inverse Autoregressive Flow models

LLM Large Language Model

LPC Linear Predictive Coding

LSTM long short-term memory

MAP Maximum Aposteriori

MFA Montreal Forced Aligner

MLLR Maximum Likelihood Linear Regresssion

MOS Mean Opinion Score

MRT Modified Rhyme Test

NSF Neural Source Filter Models

NSIM Neurogram Similarity Index Measure

OBS-P obstruent-rich phrases

OVE Orator Verbis Electris

PAT Parametric Artificial Talker

PB-50 Phonetically Balanced wordlists

PESQ Perceptual Evaluation of Speech Quality

POLQA Perceptual Objective Listening Quality Analysis

PSOLA Pitch Synchronous Overlap and Add

Q Fastpitch WaveGAN

R Tacotron WaveGAN

RTs Rhyme Tests

SAM Standard Segmental Test

SON-P sonorant-rich phrases

SPSS Statistical Parametric Speech Synthesis

TTS Text-To-Speech

VBA voice-based assistants

ViSQOL Virtual Speech Quality Objective Listener

WER Word Error Rate

Y Fastpitch WaveNet

Z Tacotron WaveNet

1 | Introduction

Human beings project human-like attributes to inanimate objects in their surroundings. Nicholas Epley theorizes anthropomorphism as a psychological tendency of humans (Epley et al., 2007). He suggests that attributing such characteristics enables us to predict the behaviour of our surroundings, and give us a survival advantage. Anthropomorphism is actively exploited in marketing strategies. Kevin the Carrot wishes us Merry Christmas at Aldi, and the mascot for Tayto crisps is a potato in a suit. Attributing human traits also stems from a compelling desire in humans to form social bonds. Adding eyes on dollar bills in a commercial ad increased humans' protectiveness towards it, and enhanced their saving behaviours (Wang et al., 2023). Anthropomorphism also drives some of the most popular inventions in the present-day AI systems. The breakthrough success of conversational AI demonstrates that talking machines can achieve high rates of acceptance in the human society.

Voice is the most natural medium of communication. So great is our reliance on voice, that we project our expectations of age, accent and gender even on robotic sounding voices (Seaborn et al., 2021; Nass et al., 1997). Moreover, we are sensitive to unnaturalness even with very minimal input (Nusbaum et al., 1997; Antons et al., 2012). Although present day voice-based agents serve a predominantly functional role, their integration with advanced conversational Large Language Models (LLMs) can result in lifelike conversational support. Agents like Alexa are social agents which have been used to help elderly patients through periods of loneliness, and have also helped to increase confidence in public speaking (Chang et al., 2018; Pradhan et al., 2019). Furthermore, just like talking to colleague is easier than e-mailing them, a voice based interaction is expected to reduce friction[1] in interacting with computers. In preparation for this possibility, we would like to establish a closer understanding of the characteristics of an agents' voice, especially with the human voice as a reference. We aim to understand how we know that a human-sounding speech is in fact real, human speech.

The main objective of this thesis is to use contrastive properties of speech segments present in corpora of synthetic speech for evaluating the human-likeness of synthetic speech. In other words, we investigate the role of the *small stuff* in speech to make us sound human.

[1] https://www.amazon.science/blog/making-alexa-more-friction-free

The BC-2013 corpus is used for our analysis, because it provides identical lexical content in older, and modern TTS voices. Features of vowels and obstruent consonants are extracted using techniques inspired from acoustic phonetics. Then, each feature is compared with the human voice to identify which features, or which segments as a whole are not produced well by TTS synthesizers. We explore whether these shortcomings, manifested as distortions, are perceivable to human listeners.

1.1 Thesis outline

Chapter 2 presents a detailed discussion on the state-of-the-art approaches in Text-to-Speech evaluation. We first provide a summarized background of techniques in TTS that will be relevant for the present thesis. Then, we provide an in-depth discussion on TTS evaluation, covering aspects of intelligibility and naturalness in older and present-day TTS synthesizers. We describe how segmental evaluation had dominated the scene of intelligibility evaluation, but has not been explored sufficiently for other perceptual attributes of TTS synthesizers. Next, we discuss the somewhat ambiguous concept of naturalness, and present a definition of naturalness as a multi-faceted perceptual attribute. Our arguments for circumscribing naturalness to human-likeness of TTS synthesizers are also presented, thus defining the scope of this thesis. Finally, we discuss how segmental evaluation can be combined with existing techniques in phonetics and speech sciences, and how it can be used to evaluate the human-likeness of TTS synthesizers.

Chapter 3 introduces the acoustic analysis framework known as the *Dive Into Divisions* approach. First, we discuss how contrastive features can be more informative than other acoustic measurements, because human listeners are more attuned to listen for them. Next, we present a complete description of the Blizzard 2013 corpus which will be used for analysis throughout this thesis. Then, we present the techniques for segmentation the corpus at a phonemic and sub-phonemic level. Particular reference to obstruent consonants is made, which require a sub-phonemic demarcation before feature extraction. This is the division part of the approach. Once marked for boundaries, the resulting segments can be analyzed through their characteristic, representative acoustic-phonetic features. These features, also called contrastive features are listed for each segment, and the extraction methods are provided. Finally, statistical analyses are conducted to identify those features or segments, that show statistically significant deviation from the human voice.

Chapter 4 extends the Dive Into Divisions approach to include modern, neural TTS synthesizers. Where only broad class categorizations were made in Chapter 3, here we systematically categorize and analyze segments in their positional, transitional and voicing related contexts. Particularly, obstruent consonants are discussed in detail. Transitional cues are captured in the features of their neighbouring vowels. Similar to Chapter 3, we conduct a statistical analysis which shows sites of segmental distortion with respect to the human voice. By categorizing segments, more diagnostic trends are revealed in terms of global acoustic consequences (such as voicelessness). This chapter demonstrates that a segmental evaluation approach can be used to identify locations of distortion in high-rated, neural TTS synthesizers as well.

Chapter 5 introduces a novel methodological framework for a subjective evaluation of TTS synthesizers. This is a specialized design suited for evaluating utterances with segmental distortion. Obstruent-rich stimuli are compared with obstruent-poor ones to investigate if their distortion is perceivable. Instead of complete utterances, stimuli of logarithmically increasing lengths are presented to the listeners. And instead of rating naturalness, participants are simply asked Does this voice sound like a human or a machine? Their binary responses are collected in a 2-AFC task. The inspiration is borrowed directly from the stimulus accumulation phenomenon, well-documented in psychophysics. which hypothesizes that a longer exposure to the stimuli should result in more accurate responses. The presentation of logarithmically increasing lengths is designed within the Weber-Fechner law of human perception.We hypothesize that a long, obstruent-rich utterance would more often be detected as machine-like. Results are analyzed through a logistic regression model, and the likelihood of accuracy is reported.

Chapter 6 provides an overarching conclusion of the methods proposed in this thesis. We discuss the limitations of our present approaches, and suggest ways of its improvement. Finally, we outline major directions of future work.

1.2 Original contributions

This thesis develops an evaluation strategy using contrastive features of phonemic units, which are described as segments. The *Dive into Divisions* approach evaluates the signal at the acoustic-phonetic level, while the *Long Arms* approach investigates the perceptual significance of the results. The original contributions can be summarized as follows:- **Chapter 2**

- Illustrated description of naturalness as a multi-faceted perceptual attribute

Chapter 3

- Introduction to the *Dive Into Divisions* approach: providing a methodological framework for analyzing segments in TTS utterances
- Embedding a core speech science perspective in TTS evaluation
  - contrastive features used as characteristic segmental features for analysis

    - corpus-phonetics techniques used for automating the analysis to handle large scale TTS output

Chapter 4

- Further categorization of segments to reveal diagnostic trends in high-quality, neural synthesizers
- Exploring the viability of neural TTS synthesizers as a tool for phonetics research

Chapter 5

- Introduction to the Long Arms approach
    - providing a novel methodological framework for subjective evaluation of TTS utterances, specifically suited for segmental distortion
    - diagnostic trends revealed between acoustic models, often overlooked in traditional MOS based evaluation
- Embedding a core behavioural science perspective in TTS evaluation utterances of logarithmically increasing lengths presented in accordance with the Weber-Fechner laws of human perception

1.2.1 List of publications

The work in this thesis has been in part disseminated in the following publications:

1. **Pandey, A.**, Edlund, J., Le Maguer, S., and Harte, N. (2023). Listener sensitivity to deviating obstruents in WaveNet. In Proceedings of InterSpeech 2023 (pp. 1080–1084). DOI: 10.21437/Interspeech.2023-1843.
2. **Pandey, A.**, Le Maguer, S., Edlund, J., and Harte, N. (2023). Natural Choice: Comparing place classification between natural and Tacotron fricatives. In Proceedings of the 20th International Congress of Phonetic Sciences, 2023, (pp. 3161-3165) ISBN: 978-80-908 114-2-3.
3. **Pandey, A.**, Le Maguer, S., Carson-Berndsen, J., and Harte, N. (2022). Formants in text-to-speech systems: Comparing TTS voices of Blizzard Challenge 2013. In Proceedings of the 33rd Swedish phonetics meeting Fonetik 2022.
4. **Pandey, A.**, Le Maguer, S., Carson-Berndsen, J., and Harte, N. (2022) Production characteristics of obstruents in WaveNet and older TTS systems. In Proceedings of Interspeech 2022, (pp. 2373-2377), DOI: 10.21437/Interspeech.2022-10606

5. **Pandey, A.**, Le Maguer, S., Carson-Berndsen, J., and Harte, N. (2021). Mind your p's and k's–Comparing obstruents across TTS voices of the Blizzard Challenge 2013. In Proceedings of the 11th ISCA Speech Synthesis Workshop (SSW 11) (pp. 166-171), DOI: 10.21437/SSW.2021-29.

All articles have been made publicly available on ResearchGate and can also be accessed through the conference website.

2 | Approaches to the evaluation of Text-to-Speech synthesizers

The evaluation of TTS synthesizers is a cross-disciplinary research field, encompassing several disciplines of signal processing, deep-learning, behavioural sciences and human-robot interaction. It is aimed at designing techniques that can accurately judge the quality of speech generated by a TTS synthesizer. The main purpose of this chapter is to provide a structured description of the prevalent techniques in TTS evaluation, and discuss their relevance with respect to the modern TTS technology. The relationship between speech science and technology is also presented, with the perspective that TTS evaluation has been an important site for reciprocal dialogue between the two communities. A progression of techniques in TTS synthesis is also provided to form the groundwork of the synthesizers used in this thesis. We discuss how each paradigm shift in TTS contributed to improvements in quality, and brought distinctive artifacts to the resultant speech.

The two major components of quality, i.e, intelligibility and naturalness are described in sequential detail. A TTS voice is considered intelligible, when there is sufficient correspondence between the intended message (i.e., input text) and the received one (i.e, the spoken form). This definition receives little disagreement. Thus, TTS intelligibility has been a primary focus of several evaluation designs, and will be described in Section 2.3. The definition of naturalness is however fraught with disagreement. Several researchers consider naturalness synonymous with human-likeness, i.e, it should be indistinguishable from a human-voice. Several other researchers argue that this definition is neither suitable nor sufficient. This is because naturalness fluctuates with how *appropriate* the voice is for a particular application or a context. Section 2.4 explains that naturalness is a multi-faceted perceptual attribute. We discuss that the human-likeness is a component of naturalness, and remains an important goal for several diverse applications of TTS synthesizers. Therefore, the central focus of this thesis is the perceived *human-likeness* of TTS voices.

Next, in Section 2.5 we discuss the historical contributions of speech science into TTS and to TTS evaluation. The relevance of phonetics in the present day TTS analysis and evaluation is also discussed. Finally, we establish that the exchange between modern phoneticians and TTS evaluation designs is focussed largely on prosodic control, and on the utterance as a whole. Segments, or smaller meaningful units of language, have received limited attention for TTS evaluation. We introduce the field of corpus-phonetics, and lay out the tools and techniques within the field which can be extended to evaluate TTS synthesizers. Finally, we present a roadmap for the thesis, drawing on the foundations discussed in the present chapter.

2.1 Text-to-Speech synthesisers

Text-to-Speech synthesizers generate speech by converting written text into the spoken form. Usually, the text is converted into an intermediate representation such as mel-spectrogram, speech parameters or phonetic transcriptions. A vocoder is then used to generate this intermediate representation into speech.

The main reference for formant synthesizers is (Klatt, 1987), while the algorithmic details have been taken from (Taylor, 2009). The structure of progression from concatenative to unit-selection synthesizers also resembles Taylor (Taylor, 2009). A thorough description of statistical parametric synthesizers, including hybrid synthesizers is presented in (Zen et al., 2009, 2013). And most recently, a comprehensive compilation of neural TTS is available in (Tan, 2023). The following sub-sections cover a summarization of TTS techniques that will be relevant for the thesis.

2.1.1 Formant synthesizers

Some of the earliest designs of an electronic formant synthesizer come from Stweart (Stewart, 1922). In this design, the lowest formants were produced through manipulating the resonance characteristics of two resonators and the source excitation was provided by a buzzer. Then in 1939, we see the appearance of Homer Dudley's “Voder” (Dudley et al., 1939), following the development of the analysis/synthesis systems in the mid 1930s, and the vocoder in 1938. The Voder was also based on manipulating source excitation through a set of bandpass filters, or resonators which mimicked the articulation mechanism of human speech production. This system involved an intricate hardware setup and was controlled by a trained human operator. The speech produced was not clear or intelligible, but it demonstrated the potential of synthesizing speech from parameters. Then, a team led by Ralph Potter designed the spectrograph, a machine that could reproduce sound into visible patterns “readily interpretable by the eye” (Potter, 1945). The broadband spectrogram, which provided a time-frequency representation of speech, contributed to several developments in speech science. For example, the observation of a “voice bar” still in practice today emerged from these developments. Following this, Alvin Liberman and his colleagues Cooper, Delattre invented the Pattern

Playback Synthesizer (Cooper et al., 1952).

The early 50s saw the the development of two important rule-based formant synthesiers: the Parametric Artificial Talker (PAT) (Lawrence, 1953) and the Orator Verbis Electris (OVE) (Fant, 1953). They were based on modelling the vocal tract as a transfer function, such that the poles of a transfer function corresponded to amplified resonant frequencies, also known as formants. The major difference between PAT and OVE was in the combination of resonators: PAT used a parallel while the OVE a series arrangement. The series type arrangement involves the transfer functions of the individual formants to be multiplied, to return an all-pole transfer function and a time domain difference equation. The series setup allows to create a transfer function for the complete vocal tract with minimal input. However, since errors can propagate through the individual formant multiplication, the synthesiser loses some controllability. Therefore, the alternative parallel setup accepts a source input individually, for each of the formants and then the resultant output is combined. The parallel arrangement was eventually favoured, and developed into multiple systems, finally leading to commercial ones. In particular, John Holmes developed a parallel formant synthesizer, where non-nasalized vowels were produced through inverse filtering of the glottal pulse (Holmes, 1973). The resultant waveform was reported to be "indistinguishable" from the human voice. This also contributed to the first implementations of KlattTalk, which was later developed as DECTalk, a commercial rule-based formant synthesizer (Klatt et al., 1984).

Although the formant synthesizers were quite customizable and had reached high intelligibility, they depended on sophisticated linguistic rules. A data-driven approach was not feasible, and that limited their progress beyond the 1980s. On the other hand, development of concatenative synthesizers (described next) was continued until recently.

2.1.2 Concatenative synthesisers

Concatenative synthesizers involved joining or concatenating segments of pre-recorded natural speech, and smoothing the resultant trajectory. Although their inception was simultaneous with formant synthesizers, their prevalence has continued until recently. Similarly, the use of LPC techniques was also influential in this concatenative synthesizers.

It was established early enough that phonemes, or syllables cannot be strung satisfactorily together, because the points of concatenation were severely influenced by their context (Harris, 1953). However, since contextual influences are minimal at the midpoint of the phoneme Peterson et al. (Peterson et al., 1958) proposed that chunks should be extracted from the acoustic midpoint of one phoneme to the other, instead of using phonemes as chunks. This unit was known as the diphone and the synthesis technique was called diphone synthesis. A large diphone inventory was required, usually of the order of the square of number of phonemes in the language. However, it provided an advantage because it reduced the reliance on specialized rules and exceptions. Additionally, contributions from the LPC techniques provided further support to the diphone synthesis approach. Linear prediction is a method to extract the filter coefficients of a signal and approximating its value as a linear sum of the values at previous timepoints (Makhoul, 1975; Markel and Gray, 1976; Taylor, 2009). For use in diphone synthesizers, the LP coefficients could be estimated from the diphones, and then joined together into an independent parameter sequence. However, the source modelling was still explicit, meaning that separate sources were used for voiced (impulse train) and unvoiced sounds (white noise). So, signal processing techniques such as the Pitch Synchronous Overlap and Add (PSOLA) were developed (Hain et al., 2005; Moulines and Verhelst, 1995). In PSOLA, the pitch and the duration of units of speech (like a diphone) are manipulated and then resynthesized to generate speech. This method greatly improves the quality of speech, and is regarded as an optimal solution for pitch and timing modification. Then, incorporating LP techniques for pitch period detection further enhanced its capabilities.

However, using pitch and timing modification was not enough, as it resulted in a number of artefacts (amplitude and stress mismatch, poor articulation of function words). The solution was offered in the form of a **unit-selection** approach, presented by Hunt and Black (Hunt and Black, 1996). This made use of larger collections of speech recordings, or corpora and thus provided a wider variability in the features of the units to be concatenated. To preserve the naturalness of the original recordings, only minimal modifications were performed on the units. The focus was instead of their selection, and their joining. These solutions were put forward in the form of costs: for selection; the target cost and for joining; the concatenation cost.

In further detail, they visualized a large-scale database of speech recordings as a “state-transition” network, where the selection of the target (state-occupancy) should minimize the target cost, and the movement to the next target (state-transition) should minimize the concatenation cost. The ideal candidate units must have minimal spectral distortion from the target units (Iwahashi et al., 1992). Also, the synthesized sequence should match the prosodic contours, required for maximizing naturalness. Usually, a pitch-correction algorithm was also applied to the generated waveform (Moulines and Charpentier, 1990).

Despite recording enormous speech databases, the complete variation required for producing speech can not be fully achieved. Hence, a distinct approach emerged in the early 2000s, known as the Statistical Parametric Speech Synthesis (SPSS). This is described next.

2.1.3 Statistical parametric synthesisers

Here, instead of selecting and concatenating units, statistical parametric synthesisers generate speech from parameters, that is feature values averaged over several instances of units. A HMM has been the most popular algorithm for modelling parameters, and later deep neural network (DNN) have also been implemented.

HMM based speech synthesisers have two primary components. The first component is the HMM, which is used to model spectral and pitch characteristics of the phonemes in a dataset, usually through a maximum likelihood estimation (MLE) approach. An HMM is created for each phoneme, which are then concatenated using decision trees for the purpose of sentence/phrase construction. Both single and multi-speaker corpora (Yamagishi et al., 2009a) can be used to develop models of the respective characteristics through HMMs. Textual features (phonological, linguistic, prosodic) are extracted from the input text. Since human speech contains both voiced and voiceless regions, a different set of parameters need to be estimated for each region. This selective pitch modelling has been explored through several methods, (Jensen et al., 1993; Ross and Ostendorf, 1994), a multi-space probability distributions (Tokuda et al., 2000) is seen as a standard technique. In the synthesis phase, a sequence of context-dependent HMMs matching the description are concatenated together. Parameters are generated for each of these HMMs in the sequence. To ensure smooth trajectories of the resultant speech, a matrix relationship between the static and dynamic parameters is also incorporated into the maximization step. Finally, a vocoder such as STRAIGHT (Kawahara, 2006), or WORLD (Morise et al., 2016) is used for the final waveform generation procedure.

A primary advantage of parametric synthesis is its adaptability to multiple speakers, accents and emotional variation. Predominantly, the Maximum Aposteriori (MAP) and the Maximum Likelihood Linear Regresssion (MLLR) techniques have been implemented to achieve adaptation in parametric synthesis. Using MLLR, speaker adaptation has been achieved with very limited amounts of speaker-specific information (Yamagishi et al., 2008; Wan et al., 2013). HMM voices are also beneficial to reduce computational footprint, because the parameters and the architecture has been shown to be stored in a few MB of space. Similarly, the robustness of HMM synthesis against recording conditions (Yanagisawa et al., 2013; Karhila et al., 2013) and data sparsity (Phung et al., 2013; Yamagishi et al., 2010) has been shown, further demonstrating its adaptable nature. However, some commonly occurring characteristics (King, 2010) of voices generated by HMMs, are creakiness, or a muffled nature. Additionally, a buzziness can be observed in the resultant speech. Often, this is attributed to spectral averaging, or oversmoothing of the speech parameters from the database.

2.1.4 Hybrid synthesisers

The hybrid approach of TTS generation combines methods from unit-selection TTS as well as HMM-based TTS techniques (Tiomkin et al., 2011). Several hybrid approaches use HMMs to guide the selection of the units (Ling and Wang, 2006, 2007). Additionally, some use units from real speech, but prosodic patterns generated by HMMs are borrowed. For example, in Plumpe et al. (Plumpe et al., 1998) an HMM was used to locate those coefficients which minimized the objective function that included parameters from both static and dynamic properties of the signal. Building from this, variable sized units (diphone, phoneme) were used for determining spectral trajectories for concatenation and the intervening dynamics respectively. On the other hand, some hybrid studies use natural speech segments interchangeably with statistically averaged speech segments (Tiomkin et al., 2011). Data sparsity also motivates the need for using these segments (Aylett and Yamagishi, 2008; Okubo et al., 2006).

Before the rise of neural TTS (described next), hybrid speech synthesis could ensure high-quality synthetic speech, as it drew advantages from both methods. Hybrid synthesizers were rated high in naturalness (King and Karaiskos, 2013; King, 2014), and were used in commercial TTS systems like Cereproc (Aylett and Yamagishi, 2008) and INNOETICS (Raptis et al., 2012, 2016).

2.1.5 Neural synthesizers

In neural synthesizers, the text-analysis module is more simplified compared to older, non-neural TTS synthesizers. Since representations can be learned directly from text, only the normalization and grapheme to phoneme conversion step is required for the pipeline. However, introducing a more granular text specification such as phrase-break prediction (Liu et al., 2020), respiratory patterns and hesitations (Yan et al., 2021; Li et al., 2023), have resulted in improvements of prosody. Similarly, semantic and syntactic representations of text embeddings have proved useful in prosodic prediction (Guo et al., 2019; Hayashi et al., 2019). The purpose of acoustic modelling is to produce features from text or phonemes. The sequence-to-sequence modelling nature of neural acoustic models overcomes several text-to-phoneme alignment steps, and is capable of producing high-dimensional mel-spectrograms as a representation of speech.

Tacotron (Wang et al., 2017) is an encoder-decoder architecture which serves as an acoustic model. A pre-net in the encoder converts text into hidden representations, which are further transformed into an output through a 1-D convolution bank + highway network + bidirectional gated recurrent unit (CBHG) module. Then, the decoder part uses the attention mechanism to the encoder input and then an RNN to generate the mel-spectrogram sequence. Although the mel-spectrogram is converted to waveform through the Griffin-Lim algorithm (Griffin and Lim, 1984) in the source text, other vocoders have also been used in the later versions of Tacotron. Additionally, the use of bidirectional LSTMs in both the encoder and decoder has further improved the resultant speech. Particularly, replacing its recurrent nature with more parallelizable architectures has gained special importance. For example, TransformerTTS can facilitate a parallel training of the encoder and decoder. And FastSpeech (Ren et al., 2020) and related models rid themselves of the attention mechanism, and instead match the duration between the phoneme sequence and the mel-spectrogram sequence.

The task of the vocoder is to convert acoustic features of a representation of speech (such as mel-spectrogram) into waveform. WaveNet (Van den Oord et al., 2016) which was the first and presently a popular vocoder was conceived at Google DeepMind. It used dilated causal convolutions which accommodates modelling of both the high-resolution and the sequential nature of speech. It outperformed the older, non-neural baselines through "never before reported" scores on naturalness, and has been highly influential in other designs of neural TTS. However, despite producing high quality speech, WaveNet struggled with slow inference speed. Therefore, several neural vocoders have been designed to improve computational speed without reducing the quality of the synthesized speech. For example, Parallel WaveNet (Oord et al., 2018) uses an Inverse Autoregressive Flow models (IAF) based model, which improves speed by predicting the samples of an utterance in parallel. Since the training of this setup is still sequential, a teacher-student setup is designed, where the student network learns from the autoregressive WaveNet through the probability distillation process. Other designs of wave generation through neural vocoders come from Generative Adversarial Assistants (GAN) based vocoders. These are composed of a generator-discriminator network, and also have a specialized loss function. In Parallel WaveGAN (Yamamoto et al., 2020) follows a standard generator-discriminator architecture, where the purpose of the generator is to deceive the discriminator between real and generated samples. However, the optimal stability of the adversarial process is achieved through a specialized loss function. This is the multi-resolution STFT loss, which offers a specialized function from multiple analysis parameters (frame shift, frame window). Another popular GAN-based vocoder is Hi-Fi GAN (Kong et al., 2020). In the generator part, a convolution network upsamples the mel-spectrogram input until the output sequence matches the desired waveform resolution. The discriminator is composed of a mixture of multi-period and multi-scale sub-discriminators, which each handle different parts of the input signals and successively evaluates audio at multiple scales. Thus, GAN based vocoders overcome the speed constraints and offer greater parallelizability in the architectures. Other methods of vocoding include diffusion based models, where random noise is first incrementally introduced into the waveform, and is reversed through a denoising procedure. The correspondence between data and the latent distribution is learnt in this way, to result in high-quality voice. Some examples are (Kong et al., 2021; Koizumi et al., 2022).

2.2 TTS evaluation: an introduction

TTS evaluation refers to the quality estimation of a synthetically produced voice, often in comparison with a previous baseline or the human voice. Although subjectively interpreted, this quality of a synthesizer can be parameterized into analyzable attributes. A semantic differentiation scaling analysis (McGee, 1964), conducted in the mid-60s revealed that "naturalness", and "intelligibility" were the two components of perceived quality of speech. Then, Klatt explains that in the time of rule-based and diphone concatenative synthesizers, evaluation was conducted on three parameters: intelligibility, naturalness and suitability to a particular application (Klatt, 1987). More recently, Hinterleitner (Hinterleitner et al., 2013) also described that "naturalness" and "intelligibility, along with "prosodic quality" were the most important components of speech quality. This means that a high quality TTS voice has predictable expectations with attributes which receive some agreement in the literature. Therefore, evaluation of TTS synthesizers has predominantly centred around these attributes. The target of intelligible speech had been achieved since the early 70s, and "never-before" (van den Oord et al., 2016) naturalness is being reported since the arrival of WaveNet. Some researchers also report that voices created by TTS are close to human-like voice in quality (Shen et al., 2018; Noah et al., 2021). Then it might appear that TTS is a solved problem[1]. Several researchers disagree.

In her seminal review of the field (Wagner et al., 2019), Wagner argues that compared to technological developments in TTS, its evaluation has not progressed with commensurate rigour. Most results are based on conventional methods, like asking a listener to *rate the naturalness* on an ordered scale. The concept of naturalness with respect to the present day synthesizers is not properly defined, and it relies on the listeners' own interpretation of the term. Just like a new disease needs new diagnostic tests, we need newer methods for detecting the weaknesses of the synthesized signal. To this end, Wagner insists that TTS evaluation should develop as its own, independent research field. Consequently, TTS evaluation has become an active area of research in the successive years. Targeted subjective tests are being designed to test synthesized speech in various contexts, lengths, conversational and interaction settings (O'Mahony et al., 2021; Clark et al., 2019; Betz et al., 2018) . Automatic methods are being developed to predict users' subjective opinion, mainly to provide feedback to the TTS engineers simultaneous with building the voice (Cooper et al., 2023; Huang et al., 2022; Hinterleitner, 2017; Lo et al., 2019). Several contributions from phoneticians are visible too: some in using the speech synthesizer as a research tool (Kirkland et al., 2022; Lameris et al., 2023; Pérez Zarazaga et al., 2023), thus pushing its limits, some in designing more diagnostic evaluation tests (Gessinger et al., 2016; Gutierrez et al., 2021).

[1] https://twitter.com/tmalsburg/status/609824961339887616
https://twitter.com/urknallen/status/1231904633653727232

Our contribution is to explore TTS evaluation through a computational phonetics perspective. Through the following discussion, we trace the historical progress of TTS evaluation in the intelligibility and naturalness of TTS synthesizers. Specifically, we highlight that segmental evaluation was an important concern for intelligibility, and also contributed to higher order attributes like naturalness. However, segmental evaluation for naturalness, or human-likeness, has received very limited attention. This can be achieved through techniques in corpus-phonetics and acoustic-phonetics, and is a major contribution of this thesis.

2.3 TTS evaluation: producing intelligible speech

2.3.1 What is intelligibility?

As mentioned in the introduction, TTS voice is considered intelligible, when there is sufficient correspondence between the intended message (i.e., input text) and the received one (i.e, the spoken form). This definition receives little disagreement. Taylor (Taylor, 2009) describes intelligibility as the “easiest problem to solve", and reports that intelligibility was already achieved in the early 70s with formant synthesizers. In the present day, intelligibility evaluation of TTS synthesizers is reported in terms of the Word Error Rate (WER) on semantically unpredictable, i.e, meaningless sentences. The most recent results on intelligibility of the state-of-the-art TTS systems report a median of 0% WER (Perrotin et al., 2023). This means that at least half of the sentences generated by most TTS systems are error-free. High-intelligibility has been achieved for low-resource languages (Xu et al., 2020; Lux et al., 2022) and found data, where transcriptions are not standardized (Baljekar and Black, 2016; Watts et al., 2013). Additionally, Cohen et al. (Cohn and Zellou, 2020) report higher intelligibility scores with older, concatenative voices. This suggests that modern, neural TTS synthesizers do not (or are not required to) bring drastic improvements to intelligibility of TTS.

Although intelligibility is not the main focus of the present thesis, early TTS evaluation had centred around intelligibility, especially at the segmental level. It was important to make the evaluation test quite diagnostic so as to pinpoint the source of distortion in the synthesized signal. The next subsection provides a detailed description of intelligibility evaluation.

2.3.2 How is intelligibility evaluated?

In this section, we review the tests and techniques used for evaluating intelligibility of TTS synthesizers. Most techniques were initially designed for intelligible communication over transmission of speech, and were quickly adopted for TTS. A wide range of tests are based on **segmental** intelligibility, that is to identify which segments (phonemes, units) in a speech stream are corrupted during transmission or synthesis. On the other hand, **sentential** tests are those that evaluate intelligibility over complete utterances, with prosodic and/or contextual support. Figure 2.1 shows the progress in these tests, concurrent with the techniques in TTS. It can be seen that segmental tests were more frequently designed, and the Modified Rhyme Tests were extremely popular. We begin with a discussion on segmental evaluation of intelligibility in TTS synthesizers.

2.3.2.1 Segmental evaluation

The Phonetically Balanced Wordlists

Among the earliest tests designed for intelligibility is the Phonetically Balanced wordlists (PB-50) (Egan, 1948), developed at the Harvard psychoacoustics lab for evaluating the intelligibility of transmitted speech signal. These wordlists were a set of 6 wordlists composed of 50 monosyllabic words each, and reflected the natural distribution of phonemes in conversational English. Moreover, in comparison to the previous tests (Fletcher and Steinberg, 1929), the evaluation procedure did not require the listener to be phonetically trained. For these reasons, these lists gained popularity in the domain evaluating the intelligibility of the transmitted speech signal, both in cases where analysis-synthesis method was used, (Bayston and Campanella, 1957; Young, 1957; David et al., 1962; Flanagan, 1960), and some other transmission techniques like time-compression (Fairbanks and Kodman Jr, 1957), or time-frequency sampling (Peterson and Subrahmanyam, 1959; Subrahmanyam and Peterson, 1959).

A common criticism of the PB-50 lists was that several low frequency words had also crept into the PB-50 list, and their rarity could alter the listeners' perception of intelligibility. The CID W-22 lists (Hirsh et al., 1952) were developed to incorporate higher frequency words which appeared in Thorndike's list of 1,000 words (Thorndike, 1932) and another list of the most frequent 128 words (Dewey, 1923) in English. Additionally, the words were presented at a homogenous volume level, as opposed to the PB-50 where this consideration was not made. Finally, the CID W-22 was a shorter 200 word list, compared to a 1000 words in the PB-50. This made the intelligibility evaluation more manageable for the listeners.

Another offshoot of the PB-50 lists appeared in the form of Consonant Nucleus Consonant (CNC) lists, developed by Lehiste and Peterson in 1959 (Lehiste and Peterson, 1959a). They argued that the phonetic balance was inadequate in the PB-50. Moreover, they pointed out that this was rather a phonemic balance, since the phonetic manifestation, or pronunciation of the phoneme was subject to too many contextual influences to be balanced. Therefore, they created 10 lists of 50 monosyllabic words that were more representative of the distribution of monosyllabic words in English. Like in CID W-22, these lists were also derived from the lists developed by Thorndike et al. However only monosyllabic words were chosen for a phonemic representation, as opposed to the PB-50 where statistics of the English corpora were used entirely.

Despite these further developments, we find that the PB-50 and its derivatives were not accepted with great ease into evaluating TTS synthesizers. For example, the acceptance of Linear Predictive Coding (LPC) techniques in speech synthesis was rapid (Atal and David, 1978; Sambur and Jayant, 1976; McGonegal et al., 1977), we see only limited evidence of PB-50 lists for their intelligibility (Chandra and Lin, 1977; Dettweiler and Hess, 1985). This means that LPC synthesizers did not use PB-50 lists very frequently for intelligibility evaluation. However, we do see PB-50 being employed for independent synthesisers used for computer aided instruction (Sanders et al., 1976) and screen reader devices (Suen and Beddoes, 1973), and their continued usage of techniques for speech transmission (Wishna, 1973; Painter et al., 1973). In the 80s, we see their application in the field of intelligibility evaluation of synthetic speech, especially for comparing commercial synthesizers like DECTalk, but were often used in combination with other evaluation tests (Pisoni et al., 1985; Greenspan et al., 1988; Nusbaum and Pisoni, 1985).

The Rhyme Tests

Following the PB-50 lists, the late 50s saw the development of a series of Rhyme Tests (RTs), initially designed by Fairbanks (Fairbanks, 1958). In these tests, evaluation was quite precise and fine-grained, because perceptual intelligibility could be evaluated at the level of a single consonantal feature at a time. These designs were particularly helpful in identifying specific locations where intelligibility of a speech communications system suffered. In the original design (Fairbanks, 1958), a recording of the stimulus word was presented to the participants. The test was a completion type, where the participants were provided with a response sheet, on which all letters except the first one were already written. For example, “-ot" was given, when the stimulus was “cot". Participants were expected to complete the spelling of the word, by providing the initial consonant. Since the initial consonant was the only variable, the exact source of confusion could be pinpointed.

However, there were two issues with this design. The first was of coverage, in that only the initial consonant was tested. And second, that the response that the listeners wrote, could be quite diverse, depending on the listeners’ subjective experience with English vocabulary (Voiers, 1967). Therefore, in later years, this design was further modified to the Diagnostic Rhyme Test (DRT) (Cohen et al., 1965), and the Modified Rhyme Test (MRT) (House et al., 1965; Griffiths, 1966). In the DRT, participants were presented with a set of minimal pairs (for example: "but" vs "cut"), each of which differed only in the initial consonant. The listener was required to select one option among the pair, reducing the variability of the responses. And in the MRT designs, consonants both in the initial and final positions were evaluated. The listeners in this case, circled their choice from a list of five alternatives.

Therefore, despite some criticism, the RTs gained much credibility as tests of intelligibility both in the domains of speech communications and synthesis. For example, for designing optimal filters (Griffiths, 1968; Arees, 1967), calculation of intelligibility scores (Voiers et al., 1965), especially in vocoded speech (Voiers, 1968; Cassel and Steele, 1963; Helms, 1968a), as well as various methods of analysis-synthesis speech synthesizers (Helms, 1968b; Strong, 1967; Carlson, 1968).

Some of the first uses of RTs to evaluate the intelligibility of synthesized speech, comes from Nye and Gaitenby in the early 70s (Nye and Gaitenby, 1973). The RTs were accepted with greater ease into the intelligibility evaluation of LPC based speech synthesisers. We find their use in diverse applications, comparing speech synthesis techniques (Keeler et al., 1974, 1976; Zahorian, 1979), and also vocoder modification techniques (Wong and Markel, 1978) for speech synthesis. Although the reason is not explicitly mentioned, we find evidence that LPC algorithms are quite sensitive to channel errors, with acceptable error rates of only 5% (Fussell et al., 1978). It is possible that the acute diagnostic nature of the RTs were a better fit for intelligibility evaluation in these cases.

This popularity of the RTs continued in later decades, especially for speech interfaces (Streeter, 1988) and TTS (Pisoni and Hunnicutt, 1980; Sherwood, 1979). Particularly, comparative evaluation of low-cost vs high-cost synthesizers was frequent, and we see active uses of the RTs in the assessment of their intelligibility (Keating et al., 1986; Greene et al., 1986; Logan et al., 1989). When HMM started to appear in the field of speech synthesis (Donovan and Woodland, 1995; Donovan, 1996; Chen et al., 1997), the RTs were borrowed into their usage as well. This means that the RTs have been an important strategy for evaluating the intelligibility of Text-to-Speech synthesizers. Figure 2.1 shows the coverage of RTs through the progression of TTS techniques. Although RTs were extremely popular, there were some studies (Greenspan et al., 1989, 1998) that pointed out that single-feature confusion (i.e, the presentation of minimal pairs) was insufficient for making a thorough enough evaluation of intelligibility. This led to the development of more specialized segmental tests, which are described next.

Other segmental tests

In addition to those pointed out (Greenspan et al., 1989, 1998), one of the primary shortcomings of such tests, was that the responses that could be elicited from the participant, were limited to the answers in the response sheet. To overcome this issue, Standard Segmental Test (SAM) (Pols and Partners, 1992), and Cluster Identification Test (CLID) (Jekosch, 1992), were developed. In SAM, syllable level stimuli of the form CV, VCV and VC were presented to

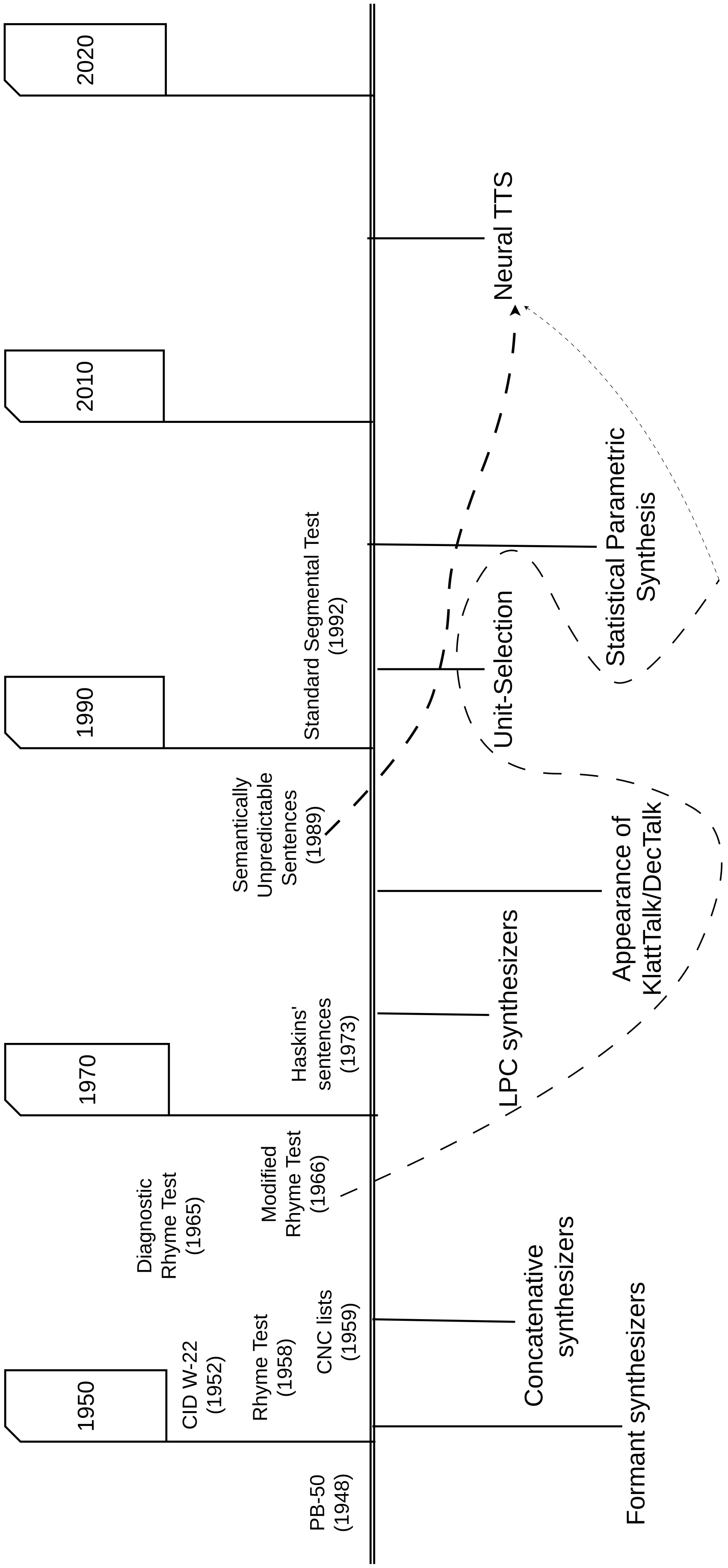

1950
1970
1990
2010
2020
PB-50 (1948)
CID W-22 (1952)
Rhyme Test (1958)
CNC lists (1959)
Diagnostic Rhyme Test (1965)
Modified Rhyme Test (1966)
Haskins' sentences (1973)
Semantically Unpredictable Sentences (1989)
Standard Segmental Test (1992)
Formant synthesizers
Concatenative synthesizers
LPC synthesizers
Appearance of KlattTalk/DecTalk
Unit-Selection
Statistical Parametric Synthesis
Neural TTS

Figure 2.1: How do we evaluate intelligibility? Segmental evaluation dominated the scene of intelligibility evaluation, but sentential tests were also in active use. Thick dashed lines trace the popularity of the evaluation test across TTS techniques. Narrowing lines in Modified Rhyme Test (MRT) indicate that although not popular, MRT is still used in some studies.

the listener, while in CLID, clusters made up the consonantal region. Participants, through this evaluation, may input whatever they perceived, instead of selecting from a limited set of responses. In addition to segmental quality, the consonant-to-vowel transitional capacity of the synthesizer could also be tested. These tests also offered syllable-level stimuli, but allowed participants to input what they heard, making their responses more open-ended. However, this setup resulted in very involved subjective listening procedures, and these tests were usually not scalable on an industrial level.

Therefore, the previous decades saw iteratively improving designs for segmental speech synthesis evaluation. However, none of these designs could faithfully diagnose users' actual experience with synthesised speech on real-world parameters.

2.3.2.2 Sentential evaluation

To overcome the problems posed by segmental evaluations, a set of sentence based evaluation tests was designed. In addition to his contribution of the PB-50 lists, Egan (Egan, 1948) also developed a collection of regular usage English sentences for evaluation of speech communication systems. This collection, known as the Harvard sentences, was a set of 68 lists of 20 sentences each. Each sentence was composed of 5 words, with 4 monosyllabic words and 1 disyllabic word. The participants were requested to write down as many words as possible, so that their correct comprehension scores indicated the intelligibility of the systems. In 1948, it was believed that sentence-level evaluation had important, but few benefits. Egan himself argued that in evaluating the intelligibility of a sentence, the listener could use cues like rhythm and meaningfulness, and gain a higher score when comprehending speech from a system (Egan, 1948; Nickerson and MILLER Jr, 1960). Despite this, in subsequent years, the use of Harvard sentences for intelligibility evaluation, can still be observed either in conjunction with the word-level tests in analysis-synthesis systems (Howard, 1956), and speech transmission systems (Kryter, 1956; Hanley, 1956), or standalone as providing test material for pattern playback synthesizers (Cooper et al., 1952). In this version, the testing format was modified further to include 72 lists of similar patterns, and the sentence length was variable (Rothauser, 1969). Following this, the Harvard sentences were used to evaluate synthetic speech produced by concatenating basic phonemes (Suen and Beddoes, 1973; Yeung, 1974), by LPC-based analysis-synthesizers (Bush, 1972), diphone based synthesizers (Wolf et al., 1978). Similarly, in subsequent years, we find the use of Harvard sentences in evaluating intelligibility of various types of speech synthesizers: concatenative synthesizers (Hauptmann, 1993; Stylianou, 1998; Sydeserff et al., 1992), commercial synthesizers (Gunderson, 1991; Venkatagiri, 1994), and later for unit-selection synthesizers (Kain and van Santen, 2007), and HMM-based speech synthesizers (Valentini-Botinhao et al., 2014; Cooke et al., 2013), and the more modern neural speech synthesiers as well (Tang, 2021; Tits et al., 2019). As an alternative format, the use of

Harvard sentences has also been extended in simply collecting user ratings on intelligibility through a mean opinion score (Stylianou, 2001; Beutnagel et al., 1998).

In addition to Harvard sentences that tested the intelligibility over proper English sentences, it was also considered important to reduce the dependence on contextual cues from the sentence. This led to the development of the Syntactically Normal Sentence Test (Nye and Gaitenby, 1973), which were syntactically proper sentences, but were not factual sentences. For example "The old farm cost the blood". These were later popularized as the Haskins sentence set. Their use was established in several studies on intelligibility of synthetic systems under development (Pisoni and Hunnicutt, 1980; Schwab et al., 1985; Jenkins and Franklin, 1982) in later years. While the Haskins sentences did reduce reliance on context, they exploited only one type of syntactic structure. The need for reducing dependence on context for intelligibility evaluation was established more rigorously, with the emergence of the Semantically Unpredictable Sentences, or the SUS (Grice, 1989; Hazan and Grice, 1989). Unlike the aforementioned sets of sentences, the SUS were not a fixed set of sentences. Instead, they were generated by grammatical rules and filled with vocabulary items, covering an exhaustive set of scenarios for evaluation. This significantly increased their flexibility, and was expanded to cover even multilingual TTS systems. (Benoit, 1990; Pols and Partners, 1992; Benoît et al., 1996). In the present two decades, the SUS have been among the most popular methodology to assess the intelligibility of synthetic systems, covering techniques like Unit-Selection (Raptis et al., 2016; Bachan and Tokarski, 2017; Cohn and Zellou, 2020) and HMM-based synthesizers (King et al., 2008; Yamagishi et al., 2009b; Picart et al., 2012; Rusko et al., 2016), and more recently neural synthesizers (Cohn and Zellou, 2020; Quintas and Trancoso, 2020).

Therefore, intelligibility of speech synthesizers was assessed both at a segmental and sentential level, where one technique complemented the other. Among these, the RTs, and the SUS continue to be used as most popular techniques for evaluation. Another important learning from this discussion, is that, since the developments in speech telecommunications technology, and speech synthesis we e often interlinked, their evaluation methodologies also saw frequent overlaps, especially upto the late 1980s. However, the development of the Haskins set of sentences and the SUS, provide a more specialized direction into evaluating intelligibility of speech synthesizers.

2.4 TTS evaluation: naturalness

2.4.1 What is naturalness?

The question of naturalness, presented usually as "how natural does this utterance sound?", has been considered "nebulous" (Wagner et al., 2019), or "poorly defined" (King, 2014), as it

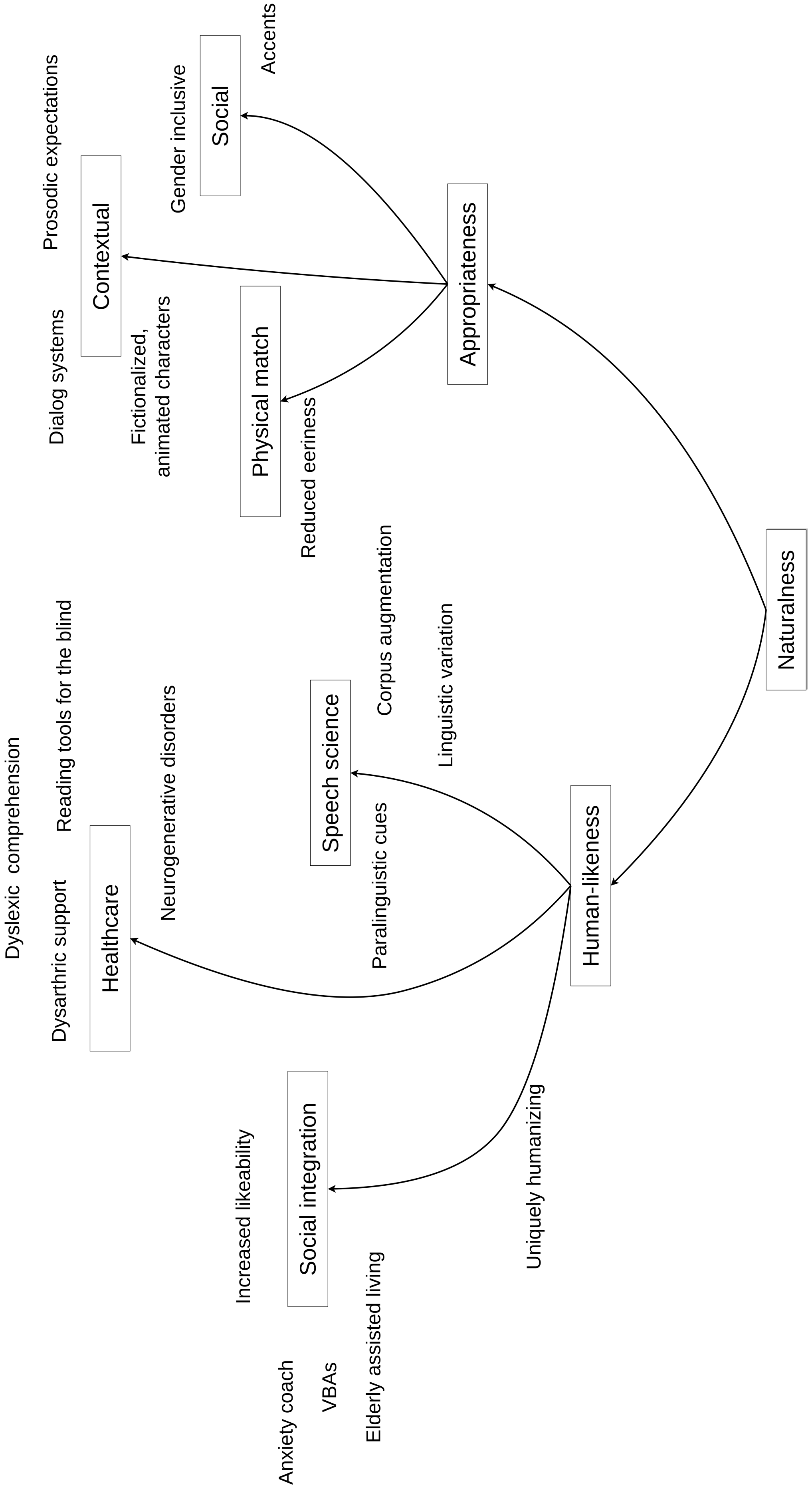


Figure 2.2: Naturalness of Text-to-Speech synthesizers a multi-component perceptual attribute. Elements of human-likeness and appropriateness both contribute to a complete social integration. Sub-components and individual applications of each component are displayed.

relies on listeners' own interpretation of naturalness. However, listeners who participate in large-scale evaluation experiment have been found to consistently agree on the evaluation results (King, 2014). In this section, we cover the extant research on naturalness in synthetic speech. Through next sub-sections, we explain that naturalness is a multi-faceted perceptual attribute, and visualize it through Figure 2.2. Human-likeness is a major component of TTS naturalness, and has diverse applications. Human-likeness of synthetic voices, or their anthropomorphism, relates to overall anthropomorphism in AI, where non-human devices are designed to represent human-like traits. However, for many applications, human-likeness in TTS neither enough, nor suitable. In such cases, it is more important for the voice to be *appropriate* to the conversation, and match the overall expectations that a user has from the conversation. We categorize naturalness into these components and the end-user applications that each delivers.

2.4.1.1 Naturalness as appropriateness

Although approaching human-likeness caters to several applications, some studies assert the importance of *appropriateness* should determine its perceived naturalness. Appropriateness can be further categorized into three aspects: contextual, physical and social, as shown in Figure 2.2. Contextual appropriateness encompasses those applications, where synthesized speech feels like a natural part of a larger discourse, or dialogue. Physical appropriateness refers to a correspondence between the physical features and its voice of a robot or an agent in a multimodal conversational setting. Finally, social contexts address the expectations from the agent in specific social settings.

First, read speech, which often serves as training data for several TTS designs is not considered appropriate in spontaneous conversational agents. For example, the MOS score for naturalness of utterances was found to vary with context-dependent instructions (Dall et al., 2014). Here, when the instruction specified placing the utterance "as part of a conversation" versus "read aloud", the rating on naturalness differed. On the other hand, when no instruction was specified, the spontaneous style was preferred. Since the experiments involved only human recorded speech, we can infer that the human-likeness of the voice alone was not the sole determinant of its naturalness. Other investigations have explored prosodic expectations of context on synthetic speech. For example, Gutierrez et al. (Gutierrez et al., 2021) situated synthetic utterances in a question-answer format (e.g. "Who ate the cake?", "MARY ate the cake.", as opposed to "Mary ate the CAKE"), and requested their listeners to mark prosodic errors. Their results showed that in sentential context influences listener expectations, such that TTS voices trained on read speech alone cannot suffice. Similarly, participants in (Wallbridge et al., 2021) used non-lexical, or discourse cues to determine those samples which prosodically matched their preceding context. These experiments, indicate that when conversational demands are explicitly specified, then contextual appropriateness is also an important part of naturalness. This holds true in human speech, and can be extrapolated to synthetic speech. Another example of contextual appropriateness comes from a corpus study on fictional characters in (Wilson and Moore, 2017). The authors show that voices of these characters revealed systematic differences from the human voice with respect to different personality types (e.g, good vs evil), and argue that the voice should fit the overall *narrative* context. In short, speech with artefacts was sometimes more important for some character personas and stories instead of human-likeness of the voice.

Similarly, the physical characteristics of the conversational agent also play an important role. Particularly, (Mitchell et al., 2011) show that a mismatch in physical features with voice can cause "eeriness". Mitchell et al. (Mitchell et al., 2011) demonstrate this through a multimodal, video analysis. Their participants accorded high levels of "eeriness" to a robot speaking in a human voice, suggesting that the mismatch was unacceptable. Similarly, McGinn et al. (McGinn and Torre, 2019) requested their participants to associate a voice with the image of a robot. The purpose here was to explore whether the participants form a mental image of the robots' physical and personality features based on its features. They found that the human voice was preferred for robots with human-like features (Nao, Stevie), while a synthetic voice was preferred for the more functional looking robots. Then, participants in Mara et al. (Mara et al., 2020) drew human-like facial features when they heard a human-like voice, and mechanical features when they heard a mechanical one. These observations appear robust across other languages, as demonstrated in Trovato et al. for Brazilian Portuguese (Trovato et al., 2015). They found that the human voice was not considered appropriate for a humanoid robot, especially when compared to a more anthropomorphic conversational agent. These findings suggest that a plain approach to human-likeness of synthetic speech is limited in an understanding of naturalness, especially in a multi-modal interaction setting.

Social appropriateness is also important for a conversation to be natural. First, a conditional, task-specific preference for robotic voices has been seen when the task is more functional. Torre et al. (Torre and Le Maguer, 2020) found that accent preferences transferred to robotic voices as well. This means that native speakers prefer an accent in robots which is already considered prestigious and trustworthy in human speech. Similarly, participants from New Zealand showed a selective preference for synthetic speech in their native accent, especially when the task involved a real-world interaction requiring trust (Tamagawa et al., 2011). Additionally, persuasive effects of native and non-native accents have been found to vary with different therapeutic approaches (Alam et al., 2021). Other aspects of social appropriateness have been associated with expectations of gender. Through a comparative analysis of several results on voice and conversational agents, Seaborn et al. observe that gender influences the perception of a robotic voice "in line with stereotypes" (Seaborn et al., 2021). For example, Nass et al. (Nass et al., 1997) report that despite explicitly removing all gender-specific information from the interaction, the perceptual attributes was stereotypical to gender roles. Similarly, a masculine voice was rated friendlier, and allowed participants to request help from the humanoid Nao (Behrens et al., 2018). Gender was also reported to interact with personality features for specific applications, such as extroverted feminine voices were better matched for healthcare agents (Tay et al., 2014). Even though the voice quality sounded non-human in older synthesizers, participants responded appropriately to gender cues (Lee et al., 2000). While it is clear that gendered expectations from robotic voices are persistent, some text-to-speech designers question whether this is *natural*. For example, methods of generating gender-neutral voices are proposed (Danielescu, 2020; Yu et al., 2022; Markopoulos et al., 2023). Although rated less natural than conventional voices (Yu et al., 2022; Markopoulos et al., 2023), Danielescu et al. (Danielescu, 2020) argue that a gender-neutral voice caters to a broader spectrum of non-binary users.

Therefore, in this subsection, we have seen that the appropriateness of the voice is important for naturalness. Listeners are sensitive to prosodic appropriateness, and a uniform training dataset is not useful. The expectation for a physical feature correspondence necessitates that there be a controllability, or a degree of human-likeness in a synthetic voice. Also, appropriateness is particularly marked for social settings, and therefore must be incorporated in TTS designs, suited to end-user applications.

2.4.1.2 Naturalness as human-likeness

In targeted applications of TTS, the concept of human-likeness is closely linked with naturalness, which is a widely tested attribute in TTS evaluation. For example, in the Blizzard Challenge series[2], where the purpose is to compare TTS techniques through a common evaluation platform, the word “natural” is implicitly used to refer to the human voice. For example, “..to find out how far our synthesizers are from natural speech." (Black and Tokuda, 2005), and “..*A* denoting natural speech" (King and Karaiskos, 2013), where *A* referred to human speech. This points to an implicit bias of “natural” = “human" in the literature. The question though is: why does a speech synthesizer need to sound like a human?

Attributing human-likeness, or anthropomorphism has historically been an important aspect of automata design (Fron and Korn, 2019). Epley et al. (Epley et al., 2007) argue that human values are also attributed to non-human, AI devices because of the compelling desire for human beings to form social bonds, especially in absence of human connections. This means that the human-likeness of automata is important for **social integration** of intelligent devices. Increasing the human-likeness of automata has been shown to increase trust (Waytz et al., 2014; Chen and Park, 2021) and error-tolerance (De Visser et al., 2016) among human users.

[2] http://festvox.org/blizzard/

Specific to the voice of automata, an important finding comes from Schroeder et al. (Schroeder and Epley, 2016), where they specifically tested the effect of a human voice on assigning human authorship to written text. Participants were asked to determine whether a given text was written by a human or AI. Text stimuli were presented to them in 4 conditions: a) text (display only text), b) audio (listen to the human speaker reading out the text), c) subtitled video (watch an actor read the text with subtitles but no audio), and d) all modalities combined (audio, visual and text). While text was generated in all cases by a text generator (Bulhak, 1996) a participant was found most likely to determine that the text was composed by a human, if the participant listened to the human voice. Further, they showed that the human voice, especially one with rich intonational variation can "uniquely humanize" an interaction, even more effectively than the audiovisual medium. The authors argue that the voice communicates the presence of a creative mind more effectively, similar to biological movements indicate presence of living beings. This work highlights the importance of the human-likeness in synthetic voices, as a primary need for human communication.

This is also supported by the many findings in voice based interaction where human, or human-like, voices are consistently rated as more socially acceptable (Schreibelmayr and Mara, 2022), pleasant (Kühne et al., 2020) and trustworthy (Weidmüller, 2022). In (Schreibelmayr and Mara, 2022), the authors tested the response of participants in to synthesized speech recordings in a variety of application specific contexts (e.g, caregiving, navigation assistance, finance assistance). Except in highly social applications like caregiving, increased human-likeness showed positive correlations with acceptance and negative correlations with eeriness. This is further supported by results on modern, neural voices (Baird et al., 2018) that show that likeability of Tacotron voices in fact, increases with their human-likeness. They argue that the uncanny valley then is not applicable for synthetic speech, at least in an audio only medium. However, a clear conception of the uncanny valley is not fully established with some disagreement within the literature (Im et al., 2023; Jansen, 2019). This thesis does not attempt to resolve the debate. Instead, we hold that human-likeness is an important target for synthesized speech.

The need for human-likeness of TTS is also explained by the socio-commercial success of voice-based assistants (VBA)s, such as Alexa, Google Home and Siri. It must be noted that their disembodied form shares no other physical features with humans, except voice. In fact, human participants of some studies have experienced doubt in their ontological classification between human and machine (Etzrodt and Engesser, 2021; Pradhan et al., 2019), and their classification as "hybrid" beings has also been reported (Weidmüller, 2022). This indicates that users interpret an interaction with VBAs socially similar to that between humans. Consequently, we find several applications of the human-likeness of VBAs for users with special needs of companionship and social assistance. For example, elderly consumers of Alexa were found to assign greater anthropomorphic characteristics to Alexa selectively during periods of loneliness (Pradhan et al., 2019). Similarly, lower levels of pitch and intonational variation was selectively preferred by participants who self-reported loneliness (Chang et al., 2018). These applications of VBAs are particularly relevant in countries where the elderly form the majority population (Chang et al., 2018). Then, participants who used Alexa as an anxiety-reducing public speaking coach suggested improvements through prosodic modulation (e.g "adding more variation in her tone while she speaks") (Wang et al., 2020a).

Human-like voice serves their classic purpose for **advancing assistive technologies and healthcare.** Text-to-speech is a crucial application for learners with difficulties, especially to promote confidence and aid in destigmatizing their educational instruction. Dyslexic children have been shown to find the human voice more intelligible (Giannouli and Banou, 2020) and has improved their comprehension on quizzes (Brunow and Cullen, 2021). Then, usability testing protocols for assistive devices developed for children with cerebral palsy also specify the requirement for a TTS to be "natural and human sounding" (Jreige et al., 2009). Developers of assistive and augmented technologies in Nepali identify that an "ideal" synthesizer must be able to "replicate speech as a human" (Basnet et al., 2023). Visually impaired listeners that used audio descriptions preferred the human voice (referred to as "natural") for dubbed feature films in Catalan (Fernández-Torné and Matamala, 2015). Similarly, in an AX discrimination task, visually impaired participants were reported to detect quality differences in synthetic voices with better precision than their sighted counterparts (Melnik-Leroy and Navickas, 2023). Through these studies, we conclude that TTS can serve as a highly enabling assistive technology.

Finally, human-like voices are required for **advancing research in speech science**. Malisz et al. (Malisz et al., 2019) argue that modern, neural synthetic speech can be a versatile tool for modelling human speech. In tasks such as lexical decision making and subjective score manipulation, participant performance and reaction times were found comparable to the human voice. If patterns, such as phonemic contrast, prosodic control and speaker variation in synthetic speech can generalize faithfully to human speech, dependence on data collection can be immensely reduced. Present studies on phonetics usually rely on carefully controlled recordings, collected at the syllable level. Although many researchers insist on using naturally occurring corpora (Chodroff, 2018; Liberman, 2019), the pre-processing of such corpora may be time-consuming as the data is often designed for other purposes. However, synthesizing high-quality, human-like speech can either augment existing corpora, or facilitate the creation of suitable corpora. Next, an important concern of speech perception is to determine the those properties of speech which remain invariant despite variation (Blumstein and Stevens, 1979). If single-speaker, syllable-level data is available as recordings (as is common), variation in positional, vocalic and cluster contexts can be synthesized, thus furthering our understanding on acoustic invariance. Additionally, the success of multi-speaker, and accented TTS (Moss et al., 2020) can ensure the necessary diversity required to understand invariant cues in speech perception. Finally, synthesizing speech can also aid in language preservation and documentation (Chasaide et al., 2015; Sakti and Nakamura, 2013).

Through this discussion, we show that human-likeness is an important target for TTS synthesizers. For targeted applications of TTS, such as healthcare, social integration of VBAs, naturalness can be synonymous with human-likeness. This thesis caters to these applications, and presents a methodology to evaluate the human-likeness TTS voices. Naturalness and human-likeness are used interchangeably to maintain consistency with the existing TTS literature. However, in Chapter 5, where we do not draw results from previous studies, the word naturalness is altogether suspended. The next section describes the various evaluation methodologies, and situates our methods within this scope.

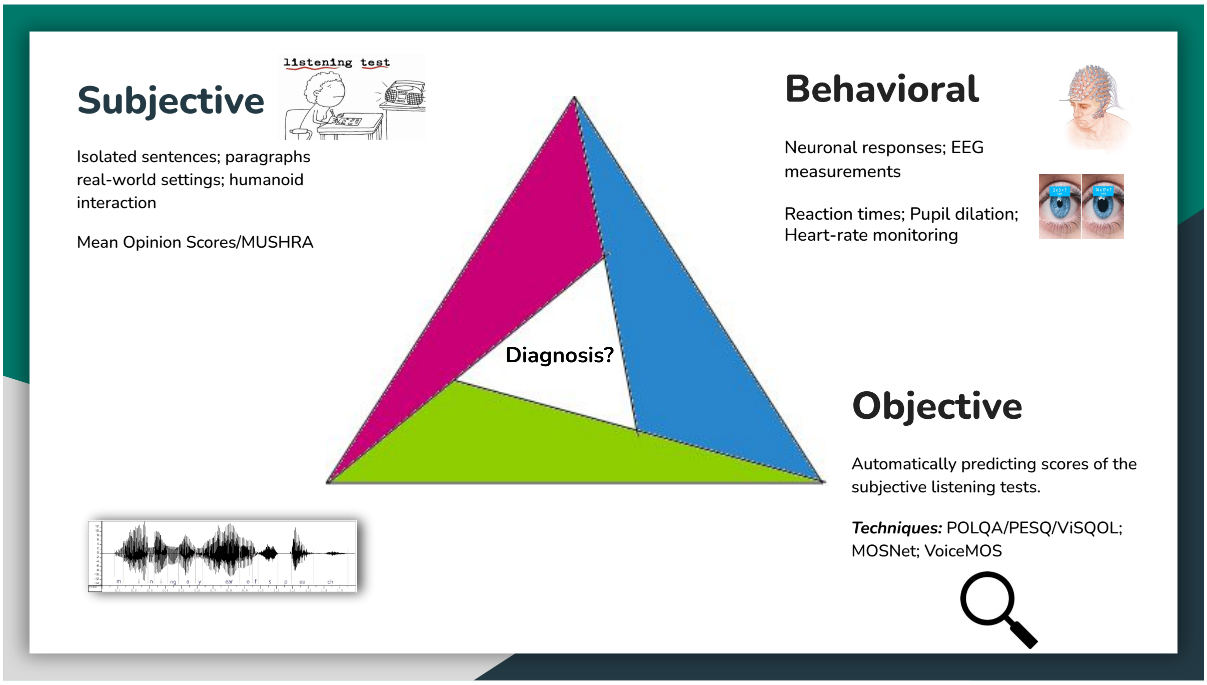
2.4.2 How is naturalness evaluated?

Naturalness evaluation of TTS synthesizers can be classified into 3 broad categories: subjective evaluation, objective evaluation and behavioural evaluation. Figure 2.3 visualizes TTS evaluation as a triangular approach. The most dominant method of evaluation is the subjective evaluation, where users' opinion is collected from them through listening tests, paired comparisons, and other interactive settings.Developing scales for such evaluation is an active resaerch area in many fields such as machine translation (Graham et al., 2013), video captioning (Graham et al., 2018) and surface realization (Mille et al., 2018). However, this approach is time consuming and expensive, so objective evaluations aim to automate this process. Finally, behavioural evaluations sidestep the collection of user opinion. Changes in physiological measures like heart rate, pupil dilation give us an estimate of the impression they had of the voice. While each framework has its own merit, a diagnostic analysis of the *signal* itself is missing from the discussion (see middle of the triangle, Figure 2.3). We describe each evaluation framework in detail. A note on segmental evaluation is also provided under subjective evaluation. Those (few) studies are highlighted where segmental evaluation has not been limited to intelligibility evaluation, but extended to attributes like system-preferences and naturalness.

2.4.2.1 Subjective evaluation: collecting user opinion

Subjective evaluation is the method of obtaining feedback directly from the consumer of the voice. Since their first appearance in the mid-60s, the MOS based design remains the dominant method in TTS evaluation until the present day in voice quality evaluation. We trace the timeline of its emergence (see Figure 2.4) , discuss its strengths and shortcomings, and analyze its alternatives.

Among several popular evaluation techniques, the isopreferent method (Munson and Karlin,



Subjective

Isolated sentences; paragraphs real-world settings; humanoid interaction

Mean Opinion Scores/MUSHRA

Behavioral

Neuronal responses; EEG measurements

Reaction times; Pupil dilation; Heart-rate monitoring

Objective

Automatically predicting scores of the subjective listening tests.

Techniques: POLQA/PESQ/ViSQOL; MOSNet; VoiceMOS

Figure 2.3: Evaluation techniques for TTS synthesizers include subjective methods, gathering listener opinions through tests; objective methods, predicting subjective scores automatically; and behavioral metrics, quantifying subconscious listener decisions and physiological changes. A *diagnosis* of the synthesized signal is missing. It can be achieved through analyzing acoustic-phonetic attributes of the TTS voice.

1962), the relative preference method and the category judgment methods (Richards and Swaffield, 1959) were among the predominant techniques. The isopreferent method involved a pairwise examination of a reference (high-fidelity signal), and a test signal. The reference signal was continually degraded, upto a point, at which the test signal and the reference signal showed equal (iso-) preference by the listeners. The relative preference method, used a test signal to be presented with a series of selected reference signals, and calculating the number of times it was preferred over the reference signals. The third and the more important method is the category judgment method, which had already received C.C.I.T.T standardization for speech quality assessment in 1962. This method is also called the Absolute Category Rating (ACR). To the best of our knowledge, some of the first applications of a mean opinion score in speech synthesis, comes from Richards et al.'s (Richards and Swaffield, 1959) design of the opinion assessment score on listening effort.

Here, listening tests involving untrained participants, were used to score the amount of listening effort that was required to understand the message that was being transmitted either as part of a conversation, or as independent sentences. Category judgments also involved planning of telephone networks, simulating a real-life conversation (Richards, 1964), and evaluating subjective variance in echo perception in a conversational setting (Richards and Buck, 1960; Williams and Moye, 1971). Participants of the experiment were required to “vote” on a 5-point scale, over which a mean was obtained. Since listening-effort scores (using a one-way setup) were relatively easier to calculate, efforts were made to connect them to conversational scores (using two-way conversation setup) as well (Richards, 1974). Another contribution of this study, was expressing subjective scores as functions of objective measurements (such as spectral information of articulation, perception, and hearing thresholds of listeners), so as to mitigate the cost of subjective listening tests .

In an important series of papers, the IBM research group presented a discussion on the comparative value of different evaluation techniques for speech quality (Pachl et al., 1968; Rothauser et al., 1971), as well as contributed one of their own designs of evaluation tests. In this design, instead of a 5-point scale of the category judgment method, participants were provided with a 10-point scale, where participants could assign even fractional scores to assess the quality of the speech signal. Another rich contribution of their work (Rothauser et al., 1971) was to provide a “preference unit (PU)” scale, i.e., a common metric, against all 4 evaluation schemes could be inter-related. Continuing in the vein, the group also used the 4 methods for a conducting a preference test over a large set of vocoded speech signals (Pachl et al., 1971). Some examples of these vocoded signals included speech samples from formant vocoders, cepstrum vocoders, rule-based, and diphone synthesizers. The relative simplicity of the category judgment method, allowed the authors to conduct an evaluation test over all the speech samples produced by the systems under comparison. However, to provide a clear ranking of systems using the combined PU scale proved very difficult, given the complexity of the testing material and several other factors.

Although no clear recommendation of a specific type of evaluation method was prescribed, the following decades saw a rise in evaluating systems using a mean opinion score, using an Absolute Category Rating method. At this stage, we have been able to identify two reasons for the popularity of MOS-based evaluation of speech. The first is that evaluation of synthesized speech can be conducted by naive, untrained listeners. This allows for large-scale comparative evaluation of different speech synthesizers, as we saw in (Pachl et al., 1971). Second, as pointed in (Grether and Stroh, 1973), ACR offered a multi-dimensional analysis of overall quality. For example, aspects like a speech signal could be evaluated along various psychological attributes like naturalness, harshness, clarity etc. As (Grether and Stroh, 1973) discuss, the pairwise comparison metrics on the other hand, made the assumption that quality was based on a unidimensional continuum.

In the mid-80s, Pols et al. (Pols and Boxelaar, 1986) describe their participation in the international ESPRIT SPIN project, which involved the development and testing of the office automation system using a speech interface. In this report, we note a few interesting points, which identify some important trends in speech synthesis. First, they make the assertion that quality and acceptability became the most important attribute of speech signals, and that systems already could provide high levels of intelligibility. Next, they propose that a MOS-based evaluation of speech synthesizers' quality, can be expressed using a multidimensional format - including criteria such as naturalness, preference, acceptability. A reference to Nusbaum et al.'s (Nusbaum et al., 1984) report is made, so it seems likely that *they* were the first to implement this. However, the exact document is not available through online/library searches. Therefore, naturalness evaluation through a MOS-based listening test was first conducted in the mid-80s.

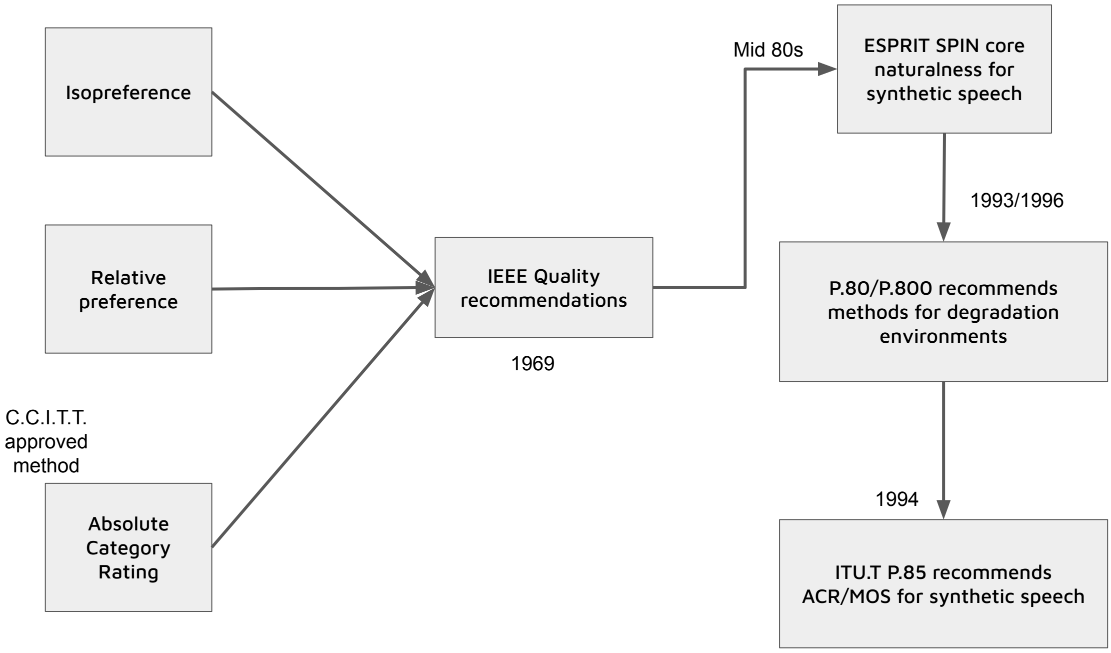

Isopreference
Relative preference
C.C.I.T.T. approved method
Absolute Category Rating
IEEE Quality recommendations
1969
Mid 80s
ESPRIT SPIN core naturalness for synthetic speech
1993/1996
P.80/P.800 recommends methods for degradation environments
1994
ITU.T P.85 recommends ACR/MOS for synthetic speech

Figure 2.4: The stages of approval of Absolute Category Judgment method through the years.

In 1992, the organization C.C.I.T.T was renamed the ITU-T [3]. Almost immediately after, in the Recommendation P.80/P.800 (Rec, 1996) [4], came the approval for using several methods for collecting the listener-opinion on quality evaluation of voice output devices. The recommendation suggested the use of these methods in any degraded listening environment, such as, echo, sidetone, environmental noise etc, and no explicit mention of speech synthesizers was made. However, in 1994, we find Recommendation P.85 (ITU, 1994) dedicated to the subjective quality testing of speech synthesizers, and the ACR, as well as Degradation Category Ratings (DCR) (Combescure et al., 1982) scales are standardized for their evaluation. Once the user inputs their ratings, a mean-opinion score is derived by taking an average of those scores. The official definition, as described in ITU.T P.10 (P.10, 2006) is, "The value on a predefined scale that a subject assigns to his opinion of the performance of the telephone transmission system used either for conversation or for listening to spoken material." Table 2.1 describes a set of

[3] source:https://bit.ly/3z0pOAJ

[4] P.80 was renumbered P.800 in 1996, and is often quoted as such

Table 2.1: Questions posed to participants evaluating different aspects of speech quality in a standard MOS based evaluation.

Aspect	Question
Overall impression	How do you rate the sound quality of the voice you have heard?
Listening effort	How would you describe the effort you were required to make in order to understand the message
Comprehension	Did you find certain words hard to understand?
Articulation	Were the sounds distinguishable?
Pronunciation	Did you notice any anomalies in pronunciation?
Speaking rate	The average speed of delivery was:-
Voice pleasantness	How would you describe the voice?
Acceptance	Do you think that this voice could be used for such an information service by telephone?

questions in a standard MOS test, when different aspects of speech quality are tested.

The subsequent years have seen active usage of MOS in assessing the quality of speech synthesizers; ranging from unit-selection synthesizers (Capes et al., 2017), HMM-based synthesizers, and finally the most modern neural synthesizers like Tacotron (Wang et al., 2017) and WaveNet (van den Oord et al., 2016). Their use has been extended to evaluation of synthetic speech in several languages (Patil et al., 2013; Kishore et al., 2003), multiple contexts, such as interactive avatar or wizard-of-oz settings (Mendelson and Aylett, 2017), crowd-sourced settings (Betz et al., 2018), as well as multi-modal speech synthesis (Mattheyses and Verhelst, 2015).

Therefore, MOS based evaluation provides subjective scores for speech quality on a discrete, ordered scale. Although these ratings give a descriptive measure of the global quality, there are some **reported issues** with the design. Some have been solved intrinsically within the paradigm of MOS, by means of a more extensive questionnaire (Polkosky and Lewis, 2003; Handley, 2009), or by re-imagining listener tests in more interactive settings (Clark et al., 2019; Mendelson and Aylett, 2017; O'Mahony et al., 2021). Le Maguer et al. (Le Maguer et al., 2024) provide a complete description of the limitations of MOS, especially asserting the misleading *relative* nature of the scale. Our main contention with MOS in this thesis is, as shown in Figure 2.3, is with its limited diagnostic abilities. We describe in Chapter 5, how new subjective evaluation methods can be designed, which can provide more diagnostic information about signal distortion.

2.4.2.2 Objective evaluations: predicting user opinion

Naturalness can also been evaluated through objective measurements. Objective methods such as algorithms or toolkits predict a measure of quality that correlates well with subjective evaluations like the MOS score. Objective methods can be full-reference or non-intrusive, dependent on the application. In the full-reference method, synthetic speech is measured against human speech as a reference, using automatic, standardized measurements like the Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001), its successors (Beerends et al., 2013) and competitors (Hines et al., 2015). Conversely, the reference signal is discarded in non-intrusive methods (Malfait et al., 2006; Kim, 2005).

The PESQ algorithm (Rix et al., 2001) follows the P.862 recommendation and is used to objectively predict the subjective quality of hand-held and telephonic devices. The perceptual model transforms both the original source signal, and the degraded signal to an internal representation, similar to the representation of auditory signals in human perception. Then, the two signals are compared sample-by-sample to predict a measure of quality of the device. Finally, a correspondence with the subjective MOS ratings is established using a correlation metric. Next, Perceptual Objective Listening Quality Analysis (POLQA) which appears as a successor to PESQ, overcomes the band and alignment limitations of PESQ. This means that, while PESQ could support narrowband and was extended to wideband signals, POLQA expanded the range to include super-wideband signals. An alternative metric known as Virtual Speech Quality Objective Listener (ViSQOL) (Hines et al., 2015) was also developed with a particular focus on degradation in voice over IP signals. ViSQOL uses a neurogram similarity measure for comparing the source, reference signal with a degraded one. A signal is divided into patches, where for every patch, a framewise comparison using the Neurogram Similarity Index Measure (NSIM) yields the point of maximum similarity between the reference and the degraded test signal. This maxima, averaged over samples returns the quality score for the utterance. The ViSQOL method provided a comparable alternative to the existing POLQA method for judging voice quality especially in VoIP settings. As can be seen, all the three methods are full-reference methods, which use the natural voice as a reference against which the transmitted, degraded or synthesized signal is compared.

Alternatively, methods that do not use a reference are known as “non-intrusive" and are considered more challenging than the full-reference ones. The standard was approved by the ITU.T P.563 (Malfait et al., 2006) recommendation in 2004, among other contenders. This method pre-processes the received signal to separate speech and non-speech regions. Then, the main sources of distortion are identified as key parameters using a set of signal parameters obtained through a series of analyses. Other designs such as ANIQUE (Kim, 2005) propose a hierarchical filterbank which models the modular decomposition of the temporal envelope in human auditory processing mechanism. Similarly, Hinterleitner et al. (Hinterleitner et al., 2010) describe a parametric approach where temporal differences between the natural reference signal and the TTS signal are quantified through an HMM trained with features of the human voice. Perceptual similarity is computed through a normalized log-likelihood score. This score, optionally combined with other features is then passed through a linear regression analysis to return a quality estimate.

Non-intrusive techniques offer reliable techniques of alleviating the dependence on the reference signal, which is often not available in real-world settings. Additionally, the mismatch between the reference human voice and that of the TTS may be enhanced due to prosodic differences. Since these metrics were mainly designed for codec conversion and impaired listening conditions, their correspondence to subjective scores on TTS is unreliable (Huang et al., 2022; Wagner et al., 2019). Therefore, many researchers are now actively exploring deep-learning based techniques further reduce the dependence on parallel natural speech as reference. An early design was the AutoMOS (Patton et al., 2016) which was developed using a family of long short-term memory (LSTM) architectures, which accepted mel-spectrograms as input. Then, fully-connected layers predicted the final MOS values through regression. When averaged over multiple utterances and speakers, AutoMOS predictions obtained a high correlation with the actual MOS scores. However, popularity of automatic MOS prediction appears to grow only 2-3 years later. The MOSNet (Lo et al., 2019) employs convolutional and Bi-LSTM architecture, while the stacking of the fully connected layers is similar to the AutoMOS. The predicted system-wide naturalness shows high correlation with human ratings. Then, prediction systems based on finetuned semi-supervised learning are designed either to capture the generalizability across listening tests (Cooper et al., 2022) or to accommodate variability in listener preferences (Tseng et al., 2021). Finally, VoiceMOS challenge conducted to compare the state-of-art techniques in objective evaluation, is now successful in its second year (Cooper et al., 2023). This further identifies an emerging trend in evaluation.

However, it must be noted that while end-to-end prediction systems automate the evaluation process, their outputs still fail to diagnose the source of distortion in the synthesized signals. Therefore, some researchers discard the Opinion Score altogether, and estimate the naturalness of the signal through behavioural and physiological metrics. The next subsection describes those methods.

2.4.2.3 Behavioural evaluations: sidestepping opinion

In addition to the inclination towards carefully designed evaluation tests, *behavioral* or *physiological* metrics have emerged. In these methods, the conscious opinion of a listener is discarded. Instead, we tap into a participants' subconscious decision making through a variety of behavioural methods. Changes in their behavioural responses - pupil dilation (Govender and King, 2018; Govender et al., 2019), neuronal activity detection (Gupta et al., 2013; Parmonangan et al., 2019; Antons et al., 2012) and reaction times can give us an estimate of the impression they have of the system.

Electroencephalography (EEG) based measurements have explored the relationship between neuronal activity in the subbands and estimation of quality in synthetic speech. An important finding comes from Antons et al. (Antons et al., 2012), who examined neuronal response to distortion in vowel level stimuli. In addition to collecting EEG data from different scalp locations, behavioral data (reaction times; onset to button click, psychometric data; how accurate as a function of SNR), and opinion test was also collected and analyzed. They found that pattern of brain activation related to processing degradations can also be detected in trials not reported as degraded. Through these methods, they identified that distortion could be detected through behavioural measures, even when these did not correspond to conscious ratings. Following these results with connected speech, Gupta et al. (Gupta et al., 2013) explored the effect of poor-quality TTS synthesizers was examined on specific event-related potentials in an oddball paradigm. Significant effects of quality were found, indicating that quality effects can be detected both with minimal exposure (vowels, (Antons et al., 2012)) and with contextual support. Another useful resource is the PhySyQx dataset (Gupta et al., 2015), where EEG data and the MOS scores are simultaneously collected. This dataset is further explored by Maki et al (Maki et al., 2018), who demonstrate that ElectroEncephaloGram (EEG) based methods can be utilized for predicting speech quality. Specifically, they identify that the neuronal activity in the alpha-band of brain-waves correlates the most with MOS prediction.

Govender et al. present a series (Govender and King, 2018; Govender et al., 2019)of pupillometry based experiments for assessing the cognitive load of quality of synthetic speech generated by various systems. In these experiments, the extent of participants' pupil dilation was measured using a pupillometer as a function of quality differences between synthesized speech. Both in semantically unpredictable and meaningful sentences, pupil dilation was found to increase as a function of quality. Additionally, recall, i.e., accurate repetition of the sentence stimuli, was also the highest in the human voice. This finding suggests that using pupillometry, the relationship between intelligibility and listening effort can be simultaneously estimated. These generalizations extended to noisy listening conditions (Govender et al., 2019).

2.4.2.4 Segmental evaluation of naturalness

Studies that discuss the contribution of segmental properties of speech in perceived naturalness of have been quite limited. However, evidence from several studies suggest that segments, such as vowels and consonants can also influence the perception of naturalness and overall quality. Segmental quality was a critical metric for overall speech quality (van Heuven and van Bezooijen, 1995), and deficiencies in synthetic voices could be detected at the word-level, even when the intelligibility was high (Klatt, 1987; Wright et al., 1986).

Evidence from electroencephalography based studies (Porbadnigk et al., 2010; Antons et al., 2012) have found that degradation in a signal can be detected by changes in neuronal activity, even when stimuli are as short as a vowel. These changes may not translate to conscious behaviour ratings, but can affect listeners' fatigue in long-term usage. Additionally, listeners' sensitivity to naturalness in synthetic speech (Nusbaum et al., 1997), has been found even at a “microscopic" level (i.e, when stimuli were only a few glottal pulses from a vowel). Studies on diphone synthesis found a significant effect of segmental features on listener preferences (Bunnell et al., 1998), and their quality was reported to influence the naturalness of intonation Vainio et al. (2002). In an investigation of unit-selection synthesisers, (Mayo et al., 2005), segmental (or unit) appropriateness was reported to be an important dimension for listeners' judgment of naturalness.

Although limited, the evidence suggests that the contribution of segmental units cannot be ignored for TTS naturalness. Techniques for a systematic analysis of segments can be borrowed directly from the field of acoustic-phonetics, where acoustic features are routinely parsed to extract the most meaningful ones from the signal. The next section describes the relevant techniques.

2.5 Speech science in TTS evaluation

2.5.1 Phonetics and speech synthesis

Phonetics is a branch of linguistics that explores the articulation, acoustics and perception of speech sounds. Articulatory phonetics aims to explain the roles of speech organs that are involved in producing various speech sounds. For example, during the production of voiceless sounds, the post-crico-arytenoid muscles are drawn inwards such that the vocal folds can be held apart with sufficient tension. Additionally, topics such as speaker variation, and speech impairment also come under the purview of articulatory phonetics. Next, comes acoustic phonetics, which provides systematic methods for analyzing the pressure variations in the air as a consequence of each of these articulation patterns. For example, a short-term Fourier transform can represent the signal into an informative time-frequency representation, from which acoustic-phonetic features can be extracted for analysis. Techniques within acoustic phonetics overlap with speech processing and technology, and are borrowed directly for the analysis of speech signals. Finally, speech perception is the branch of phonetics where we analyze which parts of the acoustic information is used by the listener to make contrastive or meaningful distinctions in language. For example, the duration of the voicing onset time is responsible for the perceived voicing difference between the English labiodental fricatives, as in *pull*: /pʰʊɭ/ and *bull* /bʊɭ/.

In the early days of speech synthesizers, speech science and synthesis technology enjoyed a reciprocal relationship. The distance between synthetic speech and phonetics grew because synthetic speech produced by early speech synthesizers was often more unintelligible than human speech, and its findings often did not extend to human language (Duffy and Pisoni, 1992). Although intelligibility improved with statistical parametric synthesis, it brought a roboticity or "unnaturalness" to the speech output (King, 2014). But today, with neural TTS, high-quality, natural-sounding synthetic speech has become quite accessible. Malisz et al. (Malisz et al., 2019) show that data-driven TTS is now more realistic, highly intelligible, and perceptually closer to human speech.

Using synthetic speech as a research tool has attracted many modern phoneticians. For example, some researchers synthesize variation in speech, in terms of disfluencies, or pause locations (Kirkland et al., 2022; Székely et al., 2019) to understand perception of paralinguistic personality traits. Similarly, architectures have been designed to support fine-grained manipulation of low-level phonetic features (Beck et al., 2022; Pérez Zarazaga et al., 2023). Also, achieving prosodic control for speaker and style transfer (Šimko et al., 2020a,b), and cuing non-explicit pragmatic functions (Lameris et al., 2023) is quite popular. In these studies, the evaluation of a TTS synthesizer is achieved rather indirectly, wherein its limits are questioned to generate nuanced speech phenomena.

By comparison, using techniques in phonetics directly for TTS evaluation has attracted modest attention. Specifically, Gutierrez et al. (Gutierrez et al., 2021) incorporate the well-established Rapid Prosody Transcription paradigm (Cole and Shattuck-Hufnagel, 2016) for obtaining locations of perceived prosodic errors. Additionally, Gessinger et al. (Gessinger et al., 2016, 2021) report differential effects of TTS techniques on phonetic entrainment patterns. These studies display that although there is increasing exchange between speech science and technology, only a few studies within phonetics target evaluation. This means that the information that resides in the acoustic-phonetic features of speech remains under-explored in TTS evaluation. The next subsection describes corpus phonetics, which is an upcoming branch in phonetics owing to open-sourced analysis and segmentation toolkits.

2.5.2 Corpus phonetics: diving into divisions

Corpus phonetics is the science of analyzing large-scale speech and language data. It derives from acoustic-phonetics, which fundamentally is the science of studying the physical, acoustic properties of speech signals. Research in acoustic-phonetics has relied significantly on speech and audio data collected at the isolated word-level, and within laboratory controlled settings. But in recent years, the availability of large-scale audio recordings (Garofolo et al., 1993; Panayotov et al., 2015; Godfrey et al., 1992), often accompanied with labeled transcription, or tools for automatic analyses (McAuliffe et al., 2017; Sonderegger and Keshet, 2012), has had an enormous impact on the field.

Additionally, in the previous two decades, research in forced alignment, both at the phonemic (McAuliffe et al., 2017) and sub-phonemic level (Sonderegger and Keshet, 2012) has produced reasonably reliable results. Similarly on the analysis front, there exist a series of toolkits that provide integrated environments for annotation, database management and statistical analysis of speech datasets.

But, as pointed out by Liberman (Liberman, 2019), the link between acoustic-phonetic properties from corpora, and their perceptual significance, is still in its nascent stages. This is because, as most speech corpora are collections of speech recordings, the analysis of acoustic-phonetic properties of speech signals is still at a descriptive level. Due to lack of accompanying perceptual data, the relationship between acoustic information in the signal, and the perceived responses is very limited.

2.5.3 Our proposal

The bottom line is that present evaluation techniques are all centred around *opinion*. Opinions in the form of user responses are collected through subjective methods, obtained physiologically in behavioural methods and predicted through objective ones. These approaches to evaluation, however, lack one thing in common: a diagnostic analysis of the TTS signal. Our proposal centres around making evaluation more diagnostic by analyzing the signal. A background in intelligibility evaluation informs us that **segmental evaluation** can pinpoint the **location** of distortion in the synthesized signal. Second, we know that the central goal in acoustic-phonetics is to connect phonetic features to meaningful, i.e, phonemic categories. In fact, a lot about *what is meaningful* is already known through decades of phonetics research. Third, we know that we can use corpus annotation and analysis techniques to analyze large scale corpora generated by TTS synthesizers. In other words, we *know* what is meaningful in the signal and we can scale it to corpora. So, we propose to conduct a **segmental evaluation of TTS corpora using corpus-phonetics and acoustic-phonetic techniques**. Using these techniques, we propose to set up a feature-by-feature comparison of natural, human speech with TTS generated corpora. We aim to establish which features differ in statistically significant ways from the human voice. Chapters 3 and 4 will be dedicated to those. However, Libermans' original issue would still remain : how does this connect to perception?

Chapter 5 is dedicated to the design of a perceptual test based in psychophysics and speech perception. Human-likeness is presented as a two-choice forced alternative (human or not). Although a Likert scale based evaluation has several advantages, as described in Section 2.4.2, a two-choice task is easier and has more inter-rater reliability compared to a Likert scale based evaluation (Awad et al., 2014). This has recently been shown to hold for neural TTS as well (Camp et al., 2023). Therefore, **paired comparisons**, already established for the quality assessment of synthetic speech (Rothauser, 1969) can particularly favour the connection between the segmental features and their perception as distinct categories. In the next section we review TTS datasets that can be utilized for such an analysis, and conclude the chapter.

2.6 A review of synthetic speech datasets

In this section, we review some of the popularly available open-source datasets for speech synthesis evaluation, and then present arguments for our selection.

First of all, the annual Blizzard Challenge series, conceptualized by Black et.al (Black and Tokuda, 2005), and maintained by King et.al [5], provide some of the most comprehensive sources of state-of-the-art speech synthesis systems. The goal of the challenge is to provide a common dataset for the participating teams, so that each team builds a synthetic voice using their own technique. All the teams are requested to generate the same set of sentences, which are then used for evaluation. The evaluation method is predominantly a subjective listening test, where the participants assign a score to the quality of the speech they hear. A MOS is calculated, and the quality of the system is thus determined. As mentioned in Section 2.5.2, participant responses, in the form of MOS, often accompany the parallel synthetic speech corpora as a final result of the challenge.

Another invaluable source of synthetic speech is the more recent Automatic Speaker Verification (ASV) Spoof Challenge series (Wu et al., 2015; Kinnunen et al., 2017; Wang et al., 2020b). While TTS technology has a great number of uses, it can also be used to impersonate someone, and misuse their biometric information. Therefore, the ASVSpoof Challenge was organized to develop counter-measures to protect a person's identity from different types of attacks. This challenge also contains a MOS-based evaluation score. However, this response is not exactly a score of system quality. Instead, it is used to measure the listener's belief on the human-ness of the speech played to them. The synthetic voices used in the dataset have high-quality, state-of-the-art voices. This makes the ASVSpoof dataset another excellent source for modeling human responses to naturalness.

Some other datasets provide various types of synthetic voices, but do not come with accompanying user responses. These include the synthetic speech commands dataset, the FoR dataset (Reimao and Tzerpos, 2019), and the SynSpeechDDB (Zhao, 2020) datasets. The first one is a single-word command dataset, while the latter two have voluminous synthetic speech data, and automatically detecting synthetic speech from real-speech. These are, however, good sources to train an automatic classifier for synthetic speech recognition.

For the purpose of our experiments so far, we have chosen the Blizzard-2013 dataset. Despite the advent of modern DNN based synthesis systems, the choice of basing our analysis on the Blizzard 2013 is motivated by multiple crucial points. Firstly, the BC-2013 provides human speech data from a single speaker, upon whose voice the other synthetic voices are modeled. Therefore, without the need for speaker normalization, we can control for variability. Secondly, the balanced audiobook rendition of the BC-2013 provided a good avenue for conducting

[5] http://festvox.org/blizzard/

fundamental acoustic-phonetic analysis, as most of the theoretical evidence is available for adult speakers.[6]. Additionally, the 300-hour training data provided with the challenge, makes it feasible for the creation and addition of state-of-the-art neural voices in future. Finally, parallel synthetic speech generated by a wide variety of techniques in BC-2013 provides room for a comprehensive comparative analysis using acoustic-phonetic attributes.

2.7 Conclusion

As modern, end-to-end speech synthesizers report near-human naturalness, better tests are required to diagnose their still existing weaknesses. The purpose of this chapter is to introduce TTS evaluation as an active and emerging research area. Evaluation tests designed for intelligibility and naturalness have been chronologically described. This perspective is presented to highlight that segmental evaluation can be quite informative, but has not been adequately investigated for naturalness. We introduce corpus phonetics, and show that its tools like forced alignment can be used to *divide* an utterance into its component segments. Acoustic phonetics can then be used to *dive* into these divisions, and analyze how (or whether) phonetic features are distorted in TTS voices.

The “nebulous" concept of naturalness has been presented as a multi-faceted perceptual attribute. Its applications as human-likeness and appropriateness are individually described. We maintain that although contextual appropriateness is necessary for a natural conversation, human-likeness remains a critical target for TTS synthesizers, with diverse applications in healthcare and social integration of robots. Moreover, since human-likeness can be encoded as a binary, categorical variable (human or not), we can connect each category to the acoustic-phonetic properties of the signal. Especially, if we find distortion in certain properties we can explore whether participants can perceive it. This is inspired from phonetics, where a central goal is to connect acoustic-phonetic features to contrast (for example, /p/-/b/ to voice onset time).

Finally, the Blizzard Challenge 2013 is used throughout this thesis. It provides data on 3 diverse TTS techniques, and now also includes modern, neural voices. The next two chapters provide a complete methodological description of our approach used on this dataset. Chapter 5 describes a novel methodology for subjective evaluation based on a paired comparison of human-likeness, based on our findings in Chapter 4. Through this thesis, we aim to make evaluation more diagnostic using techniques from phonetics.

[6] The 2014-2015 editions focused on synthesis for Indian languages, the 2016-2018 editions were based on audiobooks for children; the 2019 edition target language was Chinese.

3 | Diving into divisions: a framework for TTS evaluation

In this chapter, we introduce the Dive-into-Division approach, a methodological framework for evaluating Text-to-Speech synthesizers. This approach involves segmenting synthetic speech utterances into phoneme-level units (*divisions*) and extracting acoustic-phonetic measurements from these segments (*diving*). These measurements are then compared between human and synthesized voices. The segmentation and feature extraction techniques draw inspiration from corpus-phonetics and acoustic-phonetics. The design is situated within the larger framework of segmental evaluation of Text-to-Speech synthesizers.

3.1 Contrastive features for TTS evaluation

Native speakers of a language can differentiate between speech sounds to decode the meaning of the words of their language[1]. For example, native speakers of English can distinctly hear the difference between the words *tin* and *din*, which differ only in one sound, their initial consonant. But an equivalent difference between *teen* and *Teen*, which is clear and meaningful to a native Hindi speaker is frequently confused by English speakers. In other words, native speakers have a collective knowledge of the meaningful differences, or *contrasts* between the speech sounds of their language, and this knowledge is used to facilitate communication. An understanding of contrasts and contrastive differences can also aid the distinction between a phone and phoneme: a phone is any sound produced by human articulators, while a phoneme is a sound understood as contrastive by the native speakers of a language. It can also be useful to think of a phone as a language-independent, while a phoneme as a language-specific unit of human speech (Moore and Skidmore, 2019).

A central goal in acoustic-phonetics and speech perception is to identify those physical or acoustic properties which maximize the difference between phonemes in a language. Acoustic attributes of the signal that robustly communicate contrastive, phonemic differences in a

[1] Detailed reference (Zsiga, 2013)

native speakers' language are known as contrastive features. Babies are born with the ability to differentiate between all speech sounds. However, within months their perceptual systems attune to phonemes in the native language of their environment(s) (Kuhl, 1993; Kuhl et al., 2006). This means that their neural auditory pathways rearrange to respond to phonemic contrasts in their language. Contrastive features therefore can be understood to encode a fundamental relationship between the auditory stimuli in a child's physical environment and their responses to it. Human listeners rely on contrastive features, as is evidenced by studies in cochlear implants. In these studies, removal of acoustic cues either through noise or signal manipulation causes confusions in segmental recognition (Dorman et al., 1990; Iverson et al., 2006). In addition to perceptual importance, their acoustics also reveal articulatory characteristics of the production mechanism. For example, changes in formant values are a direct consequence of the changing shape of the oral cavity during vowel production (Zsiga, 2013; Stevens, 2000).

To the best of our knowledge, they have not been used in evaluating improperly produced segments in TTS synthesizers. Through this thesis, we introduce the segmental analysis of contrastive features for evaluating Text-to-Speech synthesizers. The approach is called the Dive into Divisions approach. First, we segment the dataset into phonemic divisions. Then, we extract contrastive features from each of these phonemes. This process is conducted over the human voice and a diverse set of TTS voices. Then, statistical models are used for a feature-by-feature comparison of the human voice with every TTS voice. We use naturally occurring, unconstrained synthetic speech corpora from the Blizzard Challenge 2013 corpus. Using computational techniques prevalent in corpus-phonetics enables automating the design. The methodological framework is detailed in Section 3.2. It describes our dataset, the BC-2013, segmentation techniques, feature extractions and experimental layouts. Then, we discuss the results in Section 3.3. We find that comparing contrastive features of segments between the human and synthetic voices, can reveal diagnostic trends, unavailable through traditional MOS-based evaluations.

3.2 Experimental framework

3.2.1 Our dataset - The Blizzard Challenge 2013

In this study, we use data from BC-2013 (King and Karaiskos, 2013). The Blizzard Challenge is an international challenge, hosted annually to compare state-of-the-art technologies in TTS. It receives widespread participation from academic and industrial institutions. Participants of the challenge are required to use the **same** training dataset to build a synthetic voice, and submit an identical set of sentences. The sentences are then evaluated using a common evaluation platform.

TTS Technique	System name	Naturalness MOS
Hidden Markov Model (HMM)	I	3.1
	C	2.9
	H	2.0
	F	1.9
	P	1.2
Unit-Selection (UnS)	L	3.0
	N	2.6
	B	2.1
Hybrid	M	3.9
	K	3.2
Human voice	A	4.8

Table 3.1: 10 TTS systems contributed by the participating teams in the original Blizzard 2013 Challenge, categorized by the TTS technique. The rightmost column displays the Mean-Opinion-Score for naturalness on a 5-point scale.

The training data in the BC-2013 was a set of audio books read by a professional voice actor. The actor was a female, native speaker of American English. The dataset was recorded by Lessac Technologies (Wilhelms-Tricarico et al., 2013) and contained up to 300 hours of audio book recordings. These were segmented only by chapters of audio books, and not by utterances. Additionally, 19 hours of utterance level speech-to-text aligned data was also released, to be optionally used by the participants of BC-2013. Specifically, Task 2013-EH1 was designed where participants were required to use the full 300-hour audio dataset.

10 teams took part in the Challenge, and submitted an identical set of 100 sentences. Of these 10 teams, 5 teams used parametric HMM or HMM-based techniques, 3 used Unit-Selection, and 2 used the Hybrid method for synthesis. Section 3.2.2 gives a brief description of these 3 predominant TTS techniques. Each team submitted a set of 100 identical sentences. Of these, 11 sentences were evaluated by 426 listeners. The sample of 11 sentences is chosen to balance the number of utterances with the number of TTS systems. A large number of listeners ensured that a proportionate distribution can been maintained in terms of expert/novice users of TTS, native speakers and non-native speakers in addition to the gender balance. Figure 3.1 provides the rankings based on naturalness.

Table 3.1 displays the MOS scores obtained by each system BC-2013, categorized according to the TTS technique. Figure 3.1 presents the global rankings of the systems in BC-2013. Figure 3.2 shows participants scoring audio samples in a traditional listening test. System A, or the human voice was given a score of 4.8 on naturalness. Among the TTS voices, system M scored the highest MOS of 3.9 on naturalness. As Figure 3.1 shows, system I, K and L received

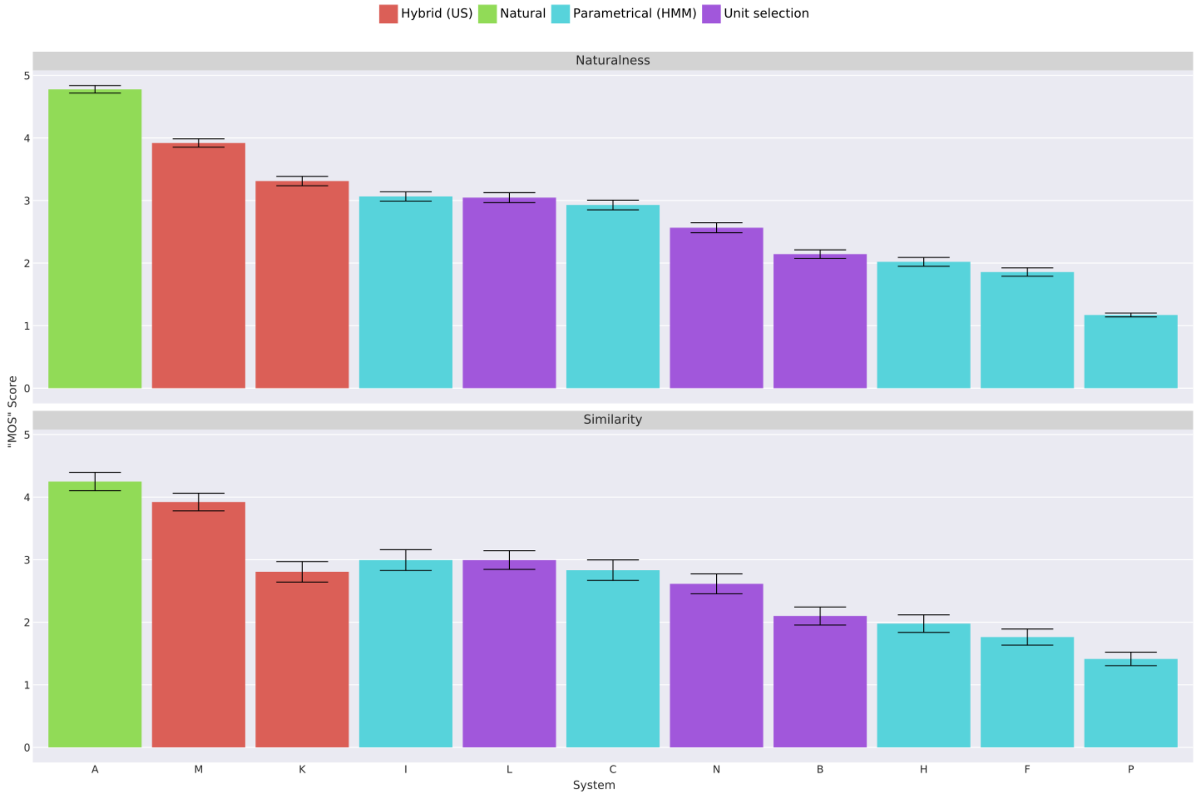


Figure 3.1: Systems of the original Blizzard Challenge 2013 ranked according to the MOS scores obtained through subjective listener tests. 10 TTS systems were built using 3 different TTS techniques. System M was ranked the highest with a MOS of 3.9, while P the lowest at 1.2. System A is the human voice.

Figure 3.2: A visual interpretation of a subjective listening test. A participant in a computer-based listening test uses headphones. Utterance level samples of synthesized and human speech are played to the participants. MOS scores are subjectively assigned to samples based on naturalness and other perceived attributes of synthesized speech like similarity, pleasantness.

a high MOS score of at least 3 points. On the other hand, the HMM system P ranked the lowest on perceived naturalness, scoring a MOS of 1.2. The next section describes the general principles of each TTS technique, and introduces how each participating team developed their own contribution to the BC-2013 Challenge.

3.2.2 TTS techniques within BC-2013

This section describes the principles of the three techniques that have been used in the BC-2013 challenge. System identifiers will not be used in the description to maintain anonymity.

3.2.2.1 Unit-selection synthesis

In the Blizzard 2013 Challenge, 3 teams submitted Unit-Selection systems. These were systems B, L and N. System B was the Festival baseline, provided by the Challenge organizers. Two teams, i.e, Innoetics/ILSP (Chalamandaris et al., 2013) and Lessaca (Wilhelms-Tricarico et al., 2013) submitted individual unit-selection systems. For selecting the target units, Innoetics combined information from the prosodic context and the part-of-speech tags to compute the target cost of the selected unit. Concatenation cost was comprised of the pitch continuity and spectral similarity. The resultant picth contour is smoothed using a polynomial interpolation method. Finally, TD-OLA performs the waveform generation. Additionally, utterances with mismatched recording conditions, alignment errors and highly deviant prosodic patterns were removed, “pruned" from the training dataset. The system submitted by Lessac technologies (Wilhelms-Tricarico et al., 2013), calculate target cost by computing the distance between acoustic variables (F0, intensity, duration) and linguistic/prosodic features encoded in a text representation. This text-based representation, known as Lessemes, predicts an intonation trajectory. Candidate units are selected based on the distance between their acoustic features and those predicted within the trajectory.

3.2.2.2 Parametric or hidden-markov-model based synthesis

In the Blizzard 2013 Challenge, 4 teams submitted HMM systems. These were systems C, H, I, P and F. System C was the HTS baseline, provided by the Challenge organizers. No documentation is available for system F. System Meraka used a cognitive theoretic model to render several emotions. The emotional state of a speaker is determined using the consequence, action and aspect of the environment. Positive or negative sentiments were then assigned to utterances through Semaffect. Semaffect could operate on a clause-level and appraise the emotional state based on the computed valence. Utterances that do not follow a conventional semantic structure are discarded, and only 30 horus of the training data is used. Quality of phonemic alignments was maintained phasewise, first at the chapter level and then realigned at the utterance and speaker level. The NITECH system used chapter adaptive training to prune out mismatched alignments between audio and phone transcriptions. Chapters are split into utterance level, which are then passed through an automatic speech recognizer HDecode (HTK version 3.4.1). Both speaker independent and speaker dependent models are sequentially used for recognition. The word error rate (WER) is calculated as the confidence measure, which is used prune out sentences from the speaker independent model.The pruned text is used to train the speaker dependent model. Moreover, a chapterwise adaptive training was used to further normalzie file formats, conditions of recording and delivery styles. While chapterwise alignment was the main contribution, it is also relevant to know that they used a multi-space probability distribution for modelling the F0 and spectrum part independently. The STRAIGHT vocoder was used for parameter generation. The Simple4All system uses freely available found data which is made more suitable with speaker diarisation. Only language-independent resources are used to minimize the dependence on expert annotation.

3.2.2.3 Hybrid synthesis

In the Blizzard 2013 Challenge, 2 systems submitted TTS systems generated using the hybrid technique. These were systems M and K.

The SHRC system (Yu et al., 2013) achieves unit-selection by maximising the likelihood of

TTS Technique	System name	Key design points
Hidden Markov Model (HMM)	C	HTS Baseline
	H	Higher order linguistic information for text analysis at front-end
	I	Chapter adaptive training for MLLR feature extraction
	P	Speaker diarization for segmenting speech into speaking styles
	F	Unavailable documentation
Unit-Selection (UnS)	B	Festival baseline
	L	Intergration of POS-tagger within the prosody model
	N	Use of existing commercial voice as suppliers
Hybrid	K	Separation of 300 hours of data into acoustic model and system building
	M	Quality data selection for alignment and audio
Natural	A	The human voice

Table 3.2: 10 TTS systems contributed by the participating teams in the original Blizzard 2013 Challenge, categorized by the TTS technique. The rightmost column displays the key design points in the development of each system.

the observation sequence, where the probability of *each* observation is predicted by context-dependent HMMs. Once the optimal sequence of units is selected, waveforms of consecutive units are concatenated to synthesize speech. Discontinuities at concatenation boundaries are smoothed with the cross-fade technique. Next, the USTC system (Chen et al., 2013) also follows a similar training and synthesis procedure. Key differences between systems are in acoustic modelling and training data selection. While USTC system uses 6 separate HMMs to model acoustic features, the SHRC uses only 2 (for spectral and excitation only). In terms of training dataset differences, SHRC uses the entire training dataset after an initial cleaning, while USTC trains their acoustic models on only 100-hours of provided data.

3.2.3 Phonemic and sub-phonemic segmentation

3.2.3.1 Phonemic segmentation

Phonemic alignments refer to the process of annotating the starting and ending point of a phoneme in the continuous spoken utterance. Employing human effort for annotation may achieve greater precision, but is quite unscalable for large corpora analysis. Therefore, speech-to-text systems have been adapted to automatically detect phoneme boundaries from continuous speech. This process is known as *forced* alingment.

We select the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). MFA uses a GMM/HMM based triphone adaptation model, with contextual support for robust, speaker-specific alignments. Each system in the BC-2013 was force-aligned separately, using a speaker-adapted triphone acoustic model pre-trained on the LibriSpeech American English corpus. To maintain consistency across pronunciations, a variant-free lexicon was created following the bootstrapping techniques in (Tanga and Bennettb, 2019; Pandey et al., 2020). Finally, 20% of these transcriptions were evaluated against gold-standard, manually annotated ones. We found that the mean error in the human voice was 7.53 ms, in the top-rated system M it was 7.1 ms, and the poorest rated systems was 8.22 ms. This means that the errors introduced by the MFA were minimal, and comparable across human and TTS voices. Therefore, in the interest of scalability of our design, we chose to retain these boundaries without further manual intervention.

Regions marked for vowels and consonants could now be extracted from the resultant phoneme boundaries. While alignment at phonemic level is sufficient for vowels and most consonants, `obstruent` consonants require a deeper, sub-phonemic level of alignment.

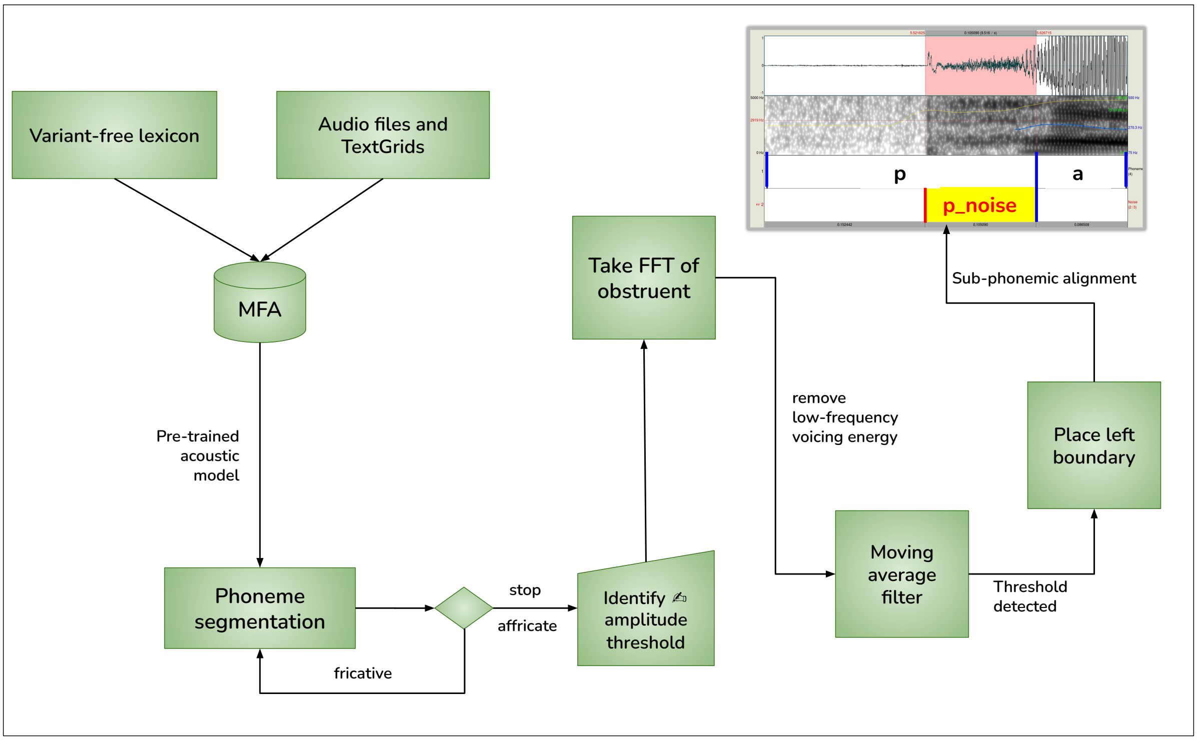


Figure 3.3: Stages in sub-phonemic alignment to separate the silence region from the noise in stops and affricates. The left boundary, i.e, the start of the noise region, is marked when the amplitude exceeds a threshold of 50-55 dB. Since noise continues until the onset of the neighbouring segment (see final step), the right boundary coincides with the end point of the phoneme, as given by the Montreal Forced Aligner (MFA). Fricatives do not require sub-phonemic alignment.

3.2.3.2 Sub-phonemic segmentation

The most important acoustic correlates of obstruent consonants are features extracted from the noisy region of the consonants. While noise continues in fricatives through the length of the consonant, in affricates and stops, it follows a region of silence. Therefore, a sub-phonemic demarcation of the noise region, separated from the silent region needed to be identified.

While most studies on obstruent contrasts depend on careful, hand-corrected methods for the analysis, it would have rendered our corpus-based approaches quite unscalable. Similarly, toolkits such as AutoVOT (Sonderegger and Keshet, 2012) require a sample of hand-annotated training data, and did not provide usable results for our corpus.

However, visually examining the spectrographic properties of stops and affricates, we found a sharp increase in amplitude, representing the burst. To extract this location automatically, we first converted the consonantal signal to its frequency domain. Then, all amplitude values <1.5 kHz were removed, because energy from the low-frequency voicing-bar interfered with the estimation of the energy of the burst. Finally, the remaining frequency-domain signal was passed through a moving-average filter. Where energy of the signal exceeded a threshold of 50-55 dB, and the point of the highest amplitude in that interval was marked as the beginning of the noise region. The threshold was decided upon after examining 20% of the sentences manually.

Algorithm 1 Marking subphonemic boundaries

```
convert signal to frequency domain
remove $< 1.5$ kHz amplitudes
apply a moving-average filter
if energy $>$ 50-55 dB then
    start point ← start of boundary at highest point
    end point ← phonemic boundary
end if
```

Thus, segmentation of each system in the BC-2013 was achieved at the phonemic and sub-phonemic level, using the MFA and the rule-based technique respectively. Representative features were extracted from each segment, as described below.

3.2.4 Feature extraction: vowels and consonants

The categorization of our segments is intentionally broad, confined to vowels and consonants alone. Further among consonants, only obstruent consonants are chosen. Firstly, because their segmentation is prone to fewer errors (DiCanio et al., 2013; Tryfou et al., 2014). And secondly, because speaker-dependent variation in nasal segments is more useful for similarity. The entire set of vowel segments, on the other hand, is used in the analysis.

3.2.4.1 Vowels: distribution and feature extraction

Ladefoged and Maddison define vowels by “the physiologic characteristic of having no obstruction in the vocal tract" (Ladefoged and Maddieson, 1996). This means that vowels are produced with an open oral tract configuration, where a constant airflow is maintained throughout the duration of the vowel. Differences in vowel production relies predominantly on the shape of the vocal tract, and are also subject to contextual variation. Production characteristics, like the position of the tongue i.e, its height and frontness-backness, and the shape of the lips (round/unrounded) are reflected in the acoustic signature of the vowel. In the BC-2013 corpus, we find instances of 12 different American English vowels /ɪ, i, æ, e, ɛ, a, ə, ɚ, ɔ, o, ʊ, u/. The chart in Figure 3.4 describes their distribution within the BC-2013 corpus.

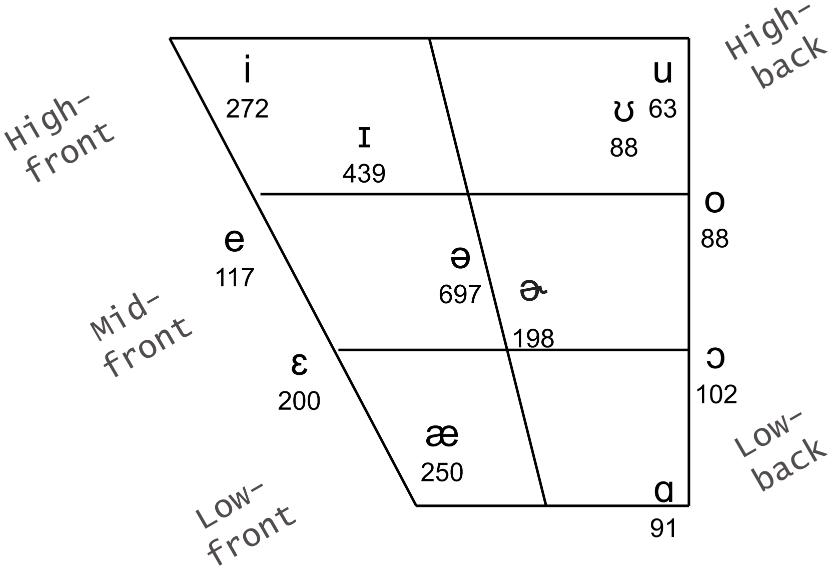


Figure 3.4: Frequency distribution of vowels in the 100 sentences of BC-2013 corpus. The height and frontness-backness of the vowel indicate the position of the tongue during vowel production. The frequency distribution is identical across systems, because every team submitted the same 100 sentences.

Vowels are characterised predominantly by their resonant properties. Vowel formants, or the peaks in the acoustic spectrum corresponding to the resonances in the vocal tract, provide important distinctive features for the perception of vowel quality. Formants can also signal a variety of paralinguistic cues when projected graphically.

Our featureset closely follows the work of Chen et al. (Chen et al., 2010) report a set of vowel space metrics for non-native speech. As in (Bradlow et al., 1996; Scarborough et al., 2007), they find that increased vowel space area and F2-F1 measurements are close correlates of intelligible speech. A brief description is provided below:-

1. **First and second formants (F1, F2)**:
   The first (F1) and second formant (F2) values are measured as the two lowest peaks in the vowel spectrum. They were extracted (Boersma and Weenink, 2018) at 20% (onset) and 50% (midpoint) of the vowels. The optimal ceiling value for each vowel was determined by the Escudero optimization procedure (Escudero et al., 2009), where the

appropriate ceiling frequency minimized the within-vowel variance in the dataset. The window size was set to 25ms. Other parameters followed the default settings in Praat.

Cross-linguistically, formants carry identifying information about the vowel identity, reflecting the tongue height and backness during vowel production. Additionally, formant transitions are considered important predictors of consonantal place of articulation (Sussman et al., 1991; McCarthy, 2019; Nearey and Shammass, 1987; Delattre et al., 1954; Liberman et al., 1954). An analysis of these informative, fundamental features is important in synthetic speech. Their malformation can cause perceptual errors and listening difficulties.

2. **Vowel space area**:
The average values of F1 and F2 from the three peripheral vowels, i.e, /i, a, o/ constitute a vowel triangle. The area of this triangle is called the vowel space area. It is defined by the following equation:-

$$area = \sqrt{s(s - D_{i,a})(s - D_{a,o})(s - D_{o,i})} \tag{1}$$

where $s = 0.5 \times (D_{i,a} + D_{a,o} + D_{o,i})$ and the euclidean distance “D" is described by:-

$$D_{x,y} = \sqrt{(F1_x - F1_y)^2 + (F2_x - F2_y)^2} \tag{2}$$

Areas of vowel spaces have shown speaker-specific differences in sex-based, age-based, pathological and emotional (Scherer et al., 2015a,b) comparisons of human speech. Specifically, reduction, or shrinkage, in vowel spaces has also been shown to accompany speech impairment (Turner et al., 1995; Shamei et al., 2023; Liu et al., 2005). If F2xF1 vowel spaces between human and synthetic speech are comparable, it may indicate proper production of vowel characteristics by the synthesizer.

3. **Overall dispersion**:
The global dispersion is calculated as the Euclidean distance of F1 and F2 values for each instance of the vowel, from the global mean of F1,F2 values for all the vowels put together. This is then averaged over the total size of the vowels in the corpus.

$$dispersion = \frac{\sum D_{i,\overline{V}} + \sum D_{a,\overline{V}} + \sum D_{o,\overline{V}}}{N} \tag{3}$$

Although only 3 vowels are displayed, dispersion was calculated for each of the 12 vowels in the corpus.

4. **Within-category dispersion**:
Here, we take the distance of each vowel from the vowel-specific mean of F1,F2 values,

instead of the global mean of vowels.

$$dispersion = \frac{1}{3} \times \frac{\sum D_{i,\overline{V_i}}}{N_i} + \frac{\sum D_{a,\overline{V_a}}}{N_a} + \frac{\sum D_{o,\overline{V_o}}}{N_o} \tag{4}$$

An itemwise statistic was also calculated by subtracting the observed formant value for every instance of the vowel with the category mean. This was useful for the linear regression analysis.

5. **Range of F1 and F2**:
   Range is calcuated as the difference between global minimum from the global maximum for F1 and F2, for every vowel.

$$R_{F1} = F1_{\text{ɑ}} - F1_{\text{i}} \tag{5}$$

$$R_{F2} = F2_{\text{ɑ}} - F2_{\text{i}} \tag{6}$$

6. **Distance between F2-F1**:
   the difference between the highest instance of F2 (for i) and the lowest F1 (for ɑ)

Area is a static measurement calculated over averaged formants, and aids visual inspection of the F2 x F1 space. The features (dispersion, ranges) offer an instance wise statistic, which are more suitable for statistical analysis of the corpora. Their extraction is described in points 3-6 above.

All 6 vowel space characteristics were extracted for each of the 10 submitted systems, as well as the human voice in BC-2013 dataset. Visual and statistical comparison between is carried out between the human voice, and each system.

3.2.4.2 Obstruents: distribution and feature extractions

Obstruent consonants are a major phonological class of consonants, accounting for 6 distinct phoneme types for stops, [p, t, k, b, d, g], 9 for fricatives, [f, v, θ, ð, s, z, ʃ, ʒ, h], and 2 for affricates [ʧ, ʤ] in English. Obstruents cover a large portion of the consonantal region in any language or dataset. Cross-linguistic evidence Lindblom and Maddieson (1988) suggests that obstruents cover between two-thirds and three-quarters of the frequency in phoneme inventories across different language groups. In the BC-2013 dataset, obstruents cover 63.9% of the total consonantal population. Their statistical dominance in the dataset makes a compelling case for their analysis. Table 3.3 presents a distribution of the obstruents in the BC-2013 corpus.

Acoustic-phonetic properties of obstruents across durational (Cho and Ladefoged, 1999; Repp, 1984; Jongman, 1989) , amplitudinal, spectral (Chodroff and Wilson, 2014; Jongman et al., 2000;

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Velar	Glottal
Stop	p b 122 130			t d 519 402		k ɡ 191 78	
Affricate					ʧ ʤ 35 32		
Fricative		f v 130 122	θ ð 53 219	s z 314 172	ʃ ʒ 96 1		h 218

Table 3.3: Frequency distribution of obstruent consonants in the 100 sentences of BC-2013 corpus. The rows represent the manners of articulation, while the columns represent the places of articulation. This distribution is identical across systems, because every team submitted the same 100 sentences.

Stevens and Blumstein, 1978) and transitional cues (Sussman et al., 1991, 1995; McCarthy, 2019) are well-established in the literature. The feature extraction procedure closely follows the methodologies presented in Jongman et al.'s seminal work on fricatives (Jongman et al., 2000), and their recent, and more comprehensive extension into all manners of obstruents (Redmon, 2020). The present discussion omits transitional cues and limits the analyses only to the consonantal portion of obstruents. The RMS amplitude has also been calculated in the frequency domain. Also, those cues which cannot be compared across all manners of articulation (for example, closure duration is only relevant for stops and affricates) are excluded.

To extract the spectral parameters, all instances of obstruents were first passed through a high-pass filter, so that the analysis spectrum remains between 550 Hz and 10,000 Hz, to separate source and filter characteristics (Shadle and Mair, 1996; Koenig et al., 2013). For fricatives, a full Hamming window was placed at the center of the frication noise. For stops and affricates, a half Hamming window was placed at the start of the burst, such that the silence region was not included. Then, spectral properties were computed using an 512-point FFT taken over these windowed signals. A brief description is provided below:-

- ***Consonant duration***: The duration of the consonantal region, as returned by the MFA. In the pre-vocalic position, this region starts with the beginning of the closure, and ends with the onset of the vowel. Conversely in the post-vocalic position, it begins at the offset of the vowel, and follows to the end of the consonant. The unit of measurement was milliseconds (ms).
- ***Noise duration***: For stops and affricates, as described above. For fricatives, since noise persists through the length of consonant, the entire region was included. The unit of measurement was milliseconds (ms).
- ***RMS amplitude***: The root-mean-squared amplitude of the power spectrum.
- ***Peak amplitude***: The value of the highest amplitude in the spectrum. The unit of

measurement is dB.

- ***Peak frequency***: This is the spectral frequency at which peak amplitude was identified. Its value was measured in Hz.
- ***Dynamic amplitude***: The difference between the peak amplitude, and the minimum amplitude below 2 kHz. The unit of measurement was dB.
- ***Spectral tilt***: The frequency domain of the spectrum was log-transformed, and then a least-squares regression line was fitted through it. The slope of this line returned the spectral tilt.

These features were extracted for obstruent consonants across all the systems, as well as the human voice, independently. The purpose of such an extraction was to compare these features across all the systems, and to identify those features, where the system (or groups of systems, See Section Table 3.6) showed significant differences from the human voice.

3.2.5 Statistical models

3.2.5.1 Spearman's correlation

The Spearman correlation measures the strength of ranked correlation between two variables. It is described by the following equation:-

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{7}$$

In this equation, ρ or r represents the Spearman's rank correlation coefficient. d$_i$ represents the differences between the ranks of corresponding data points in the two variables, and n is the number of data points. The value of ρ can range between -1 and +1. A high positive value indicates perfectly correlated variables, where a value of 0 indicates no correlation. We begin with the null hypothesis that there will be no correlation between the calculated feature means and the collected MOS scores. The significance of the correlation test is calculated on the basis of its t value and is given by the following formulation:-

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \tag{8}$$

The corresponding statistical significance is determined against the t distribution table of n-2 degrees of freedom. If this value is less than the one given for the value at df = n-2, then the null hypothesis is rejected.

For our experiments, the Spearmans' rank correlation was used to establish the relationship between the segmental features and the MOS scores on naturalness. As discussed in Section 3.2.1, MOS captures naturalness in a single, per-system value, averaged over several participants and utterances. Therefore, feature values were also computed as a single statistic per system.

For example, if MOS for every system in the BC-2013 is given by a vector $\text{MOS}_{BC-2013}$ = [MOS_M, MOS_K, ..., MOS_P], where the subscript represents the name of each system. Vowel space area computed over the entire set of vowels can be described as: $\text{Area}_{BC-2013}$ = [Area_M, Area_K, ..., Area_P]. If a high positive correlation was found, it indicated that increasing vowel space area increases the perceived naturalness. This helped us connect the predictive capacity of each acoustic-phonetic measurement with the perceived scores.

3.2.5.2 The linear mixed-effects regression model

A linear regression analysis models the relationship between independent and dependent variable. In its simplest form, it is described by the following equation:-

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \tag{9}$$

Here, y is the dependent variable, and the x_1 the independent variable. The task of modelling is to estimate the coefficient β_1, which describes the relationship between x and y. A positive value indicates a direct relationship, while a negative an inverse one between the two variables. When no influence of the independent variable is found i.e, β_1 is 0, then the dependent variable is given by the intercept β_0. Finally, ε is the error term which accounts for the uncertainty in the model, particularly when the model is not a perfect fit to naturally occurring, real-world data.

The influence of multiple independent variables can also be examined on y, where the equation can take the following form:-

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon \tag{10}$$

These variables are also known as fixed effects. A model created with multiple fixed effects is a linear *mixed* effects model. A random effect can also be added, in case observations of a certain sub-group in a population are expected to exhibit similar characteristics.

In analyzing the features of both vowels and obstruent consonants, the feature value is always the dependent variable, the human voice the intercept and the TTS technique (or individual system) is the dependent variable. An example model equation is:

$$Vowel.Dispersion = Human + \beta_1\, TTS.Technique + (1|Utterance) + \varepsilon \tag{11}$$

Through this model, we examine *how* each TTS technique influences the value of a segmental feature, in comparison to the human voice. Notice that the feature value calculated for the human voice was considered the reference point (the intercept) in each case. The **deviation** from this voice was the comparative metric across which different behaviours of groups were recorded.

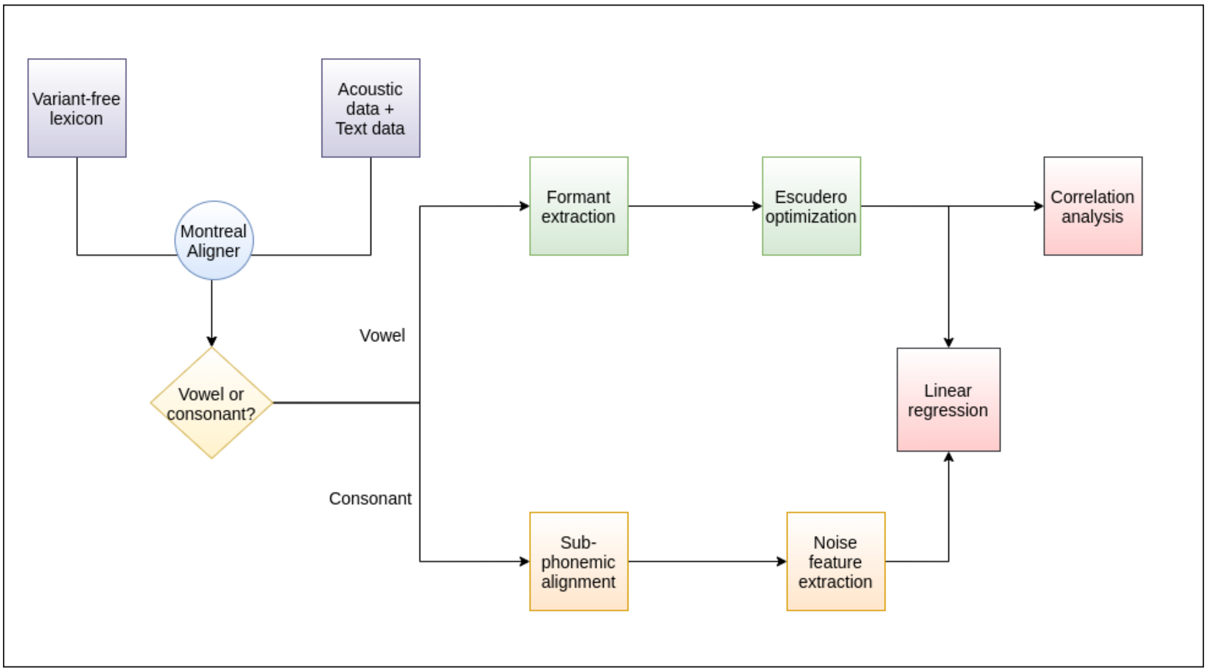


Figure 3.5: Pipeline describing the various stages of forced alignment, feature extraction and statistical modeling. The type of analysis depends on the nature of the phonemic segment.

3.2.6 Experimental layout: vowels and obstruent consonants

Contrastive features in vowels and consonants provide a range of features for analysis. In each experiment, we aim to compare each feature between the human voice and every synthetic voice in the BC-2013. Since the focus of this chapter was to build the methodology, only the voice of a single, female speaker has been used as the reference voice throughout. (See Section 3.2.1 for speaker details) A method for its extension to multiple speakers has been discussed in Chapter 6, Section 6.2. Segmentation, feature extraction and statistical comparison between the human and synthetic voices are common in both vowels and consonantal analyses. However, their experimental designs differ in the following ways.

First, vowel features are analyzed on the basis of a correlational model. Average feature values per system have been correlated with the MOS scores obtained per system. Features found significant are explored in further detail, using the linear mixed-effects regression model. But for consonants, we did not find significant correlations between the MOS and individual features of obstruent consonants. This means that, segmental features could not be globally correlated with MOS. Therefore, we localized the analysis to a within-technique one. Systems of the same TTS technique (e.g. HMM, Unit-Selection) were *grouped* and compared only for quality differences. Since these results were found statistically important, they are presented in detail. Figure 3.5 displays the shows the basic workflow of the experimental pipeline.

3.3 Results

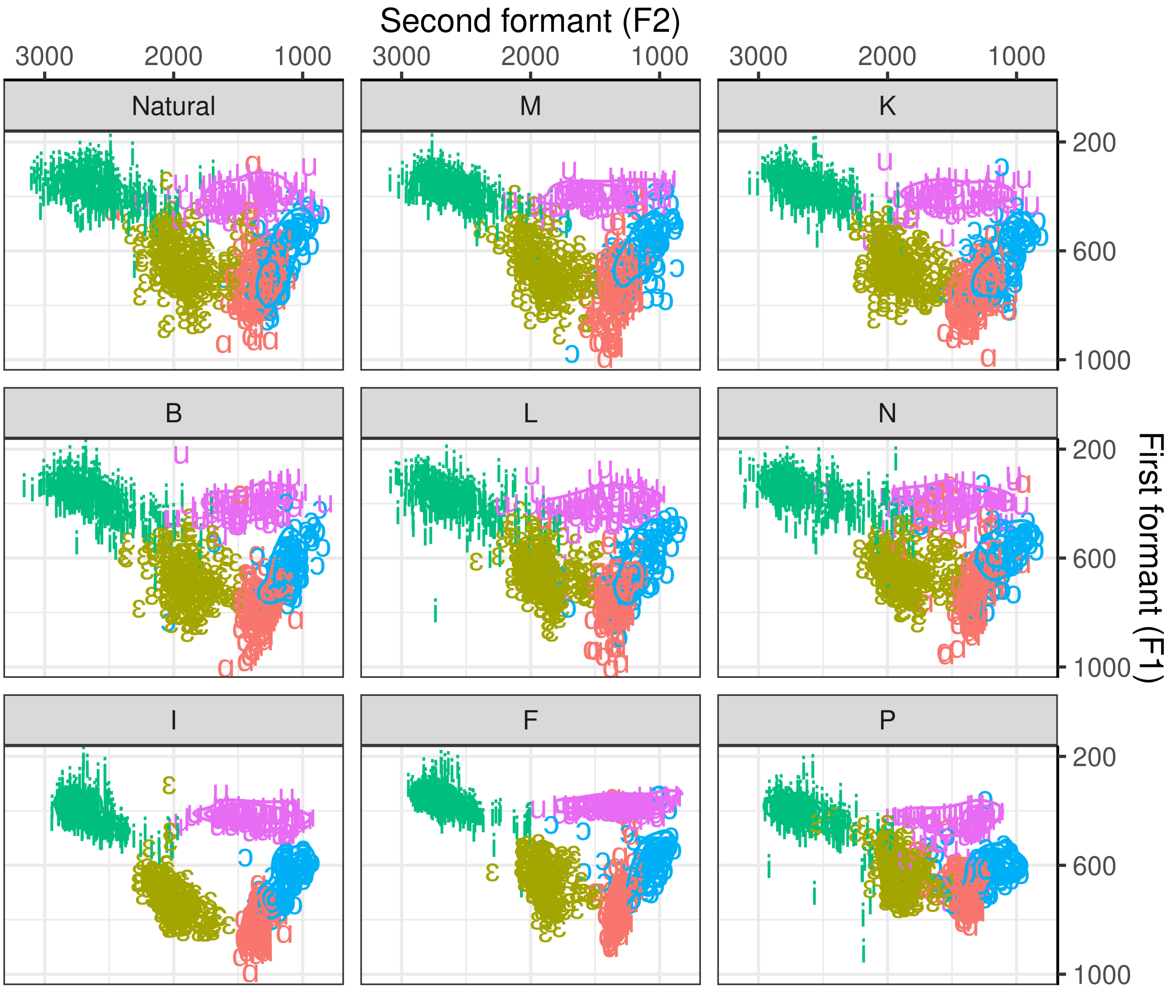


Figure 3.6: Vowel formant plots for Systems A-P, arranged in rows by TTS-technique, and in columns by quality. The first row has Natural, Hybrid systems M, and K. The second row has 3 unit-selection systems (L, N, B) And the last two rows have HMM systems I, C, H, F and P. HMM systems can be clearly seen to cluster instances of vowels around respective vowel means.

3.3.1 Analysis of vowels

3.3.1.1 Visual analysis of the vowel space area

Figure 3.6 displays the F2xF1 vowel space area plots for the systems in the BC-2013 dataset. These measurements are displayed for 5 vowels /ɛ, ɔ, i, a, u/, which had maximum separability in the vowel space.

The human voice exhibits a wide area, where instances of each vowel are dispersed around the mean. Hybrid and unit-selection systems also replicate this pattern well. This indicates that vowel units, both in the hybrid and unit-selection systems retain variability in vowel instances that appears similar to human speech. By contrast, HMM synthesis exhibits a **more contracted** vowel space. Vowel instances in systems I-P, regardless of quality can be seen contracted within their categories, reducing the vowel space area. Specifically, system P appears to centralize all of its vowels.

Figure 3.6 also shows that low back vowel a exhibits appears to have merged with the central back vowel ɔ. In all the other systems, however, these vowel categories are more distinct. These observations are visually informative, and motivate the need for a quantitative analysis.

3.3.1.2 Correlation model: which vowel features correlate with MOS?

Table 3.4 displays the correlation with each of the features for a p-value < 0.05. A positive correlation means that as the value of the concerned characteristic increases, so does the MOS score on naturalness.

Within-category dispersion in vowels was found to have a positive and modestly significant correlation with MOS on naturalness [r(8) 0.64, p-val ≤ 0.05]. Dispersion values in humans voice is recorded at about 742.59 Hz. In Hybrid and Unit-selection systems, the dispersion values are comparable to the human voice, within a 650-700 Hz range. But in HMM systems, this value shrinks to 500-550 Hz. Therefore, we see that the correlation is also motivated by the TTS technique. Rankings between individual systems of the same technique also exhibit this trend. The higher ranked hybrid system M shows a wider vowel-specific space, than lower ranked K. Similar comparisons can be made between unit-selection systems, L and B. However, naturalness ratings in HMM systems do not correlate with within-category vowel dispersion. Regardless of ranking, HMM systems tightly cluster the vowel instances around their category means.

Next, **F1-range** was also found to correlate with the perceived MOS, with statistically significant effects [r(8) 0.73, p-val < 0.05]. This vowel-independent F1 range for the human voice is about a 1087.18 Hz. The highest ranked system M comes closest to this value, at 1066.05 Hz. This value does not exceed 880 Hz in almost all other systems BC-2013 systems, regardless of

Feature	Correlation with MOS
Area	0.05
Dispersion	0.24
W-C Dispersion	**0.64**
F1 range	**0.73**
F2 range	0.55
F2-F1 (aa)	-0.07
F2-F1 (iy)	0.07

Table 3.4: Spearman's correlation of each feature with MOS ratings for similarity and naturalness. Values in bold denote significant correlations determined using the t-distribution at df=n-2. Dataset comprised of a total 2605 vowels (individual descriptions in Figure 3.4) from 100 test sentences of Blizzard Challenge 2013.

the TTS technique. This shows that a broad range of F1 values can be perceivable as greater naturalness.

3.3.1.3 Statistical model: which TTS techniques affect features?

As can be seen in the above description, the within-category dispersion was most predictive of naturalness. Therefore, we analyzed this feature using a linear regression model (Kuznetsova et al., 2017). The system type (*HMM*, *Hybrid* and *Unit-Selection*) as well as the vowel type (*front* or *back*) are considered as fixed effects. The effect of the utterance is coded as a random effect shown as $(1|\mathit{Utterance})$. This is because the lexical content of individual utterance (e.g, narration style, position in the audiobook chapter) is assumed to not influence the TTS generation techniques.

$$\mathit{Dispersion}_i \sim \mathit{TTS.technique} + \mathit{Vowel.Type} + (1|\mathit{Utterance}) \tag{12}$$

The results are presented in Table 3.5.

The results demonstrate that the human speech produces more dispersed F1 values than any type of synthetic speech. In addition, the dispersion of F1 values for HMM synthesis is significantly less than the other types of synthesis. HMM synthesis produces even less dispersed F2 values. These results support the trends observed in the visual inspection and the correlational analysis. We see that HMM synthesis produces more distinct vowels than any type of synthesis, but it fails to generate more extreme or more nuanced vowel instances which may be fundamental to naturalness. This is a consequence of the well-known over-smoothing effect, due to the statistical nature of parametrical synthesis (Zen et al., 2009), which leads to a reduction of the variability of the generated speech.

Additionally, we observed a significant effect of vowel-type on dispersion, where front vowels were seen to lower dispersion (β_{front} = -5.933, 95% CI [-9.71, -2.16], p $<$ 0.001). According to

	System	β	95% CI	p <
F1	Unit Sel	-7.38	[-11.44, -3.31]	0.01
	Hybrid	-11.58	[-15.90, -7.26]	0.001
	HMM	-29.12	[-33.06, -25.18]	0.001
F2	Unit Sel	-7.49	[-13.80, -1.17]	0.01
	Hybrid	-9.80	[-16.50, -3.10]	0.05
	HMM	-40.18	[-46.29, -34.06]	0.001

Table 3.5: F1 and F2 dispersion analysis

the degree of articulatory constraints (Recasens et al., 1997), front vowels, /i/ is quite resilient to coarticulatory influence from surrounding contexts. The lack of dispersion in the front vowels, could be contributed to the frequent presence of this vowel. However, to investigate an interaction between the system-type and vowel-type, based on their indivdual coarticulatory resistance, would be quite interesting as future work.

3.3.1.4 Vowels: summary of results

In this experiment, we compared representative features of vowels between the human voice and the synthetic voices of the BC-2013. We found that the correlation between within-category vowel dispersion patterns and the Blizzard Challenge 2013 MOS scores for naturalness were statistically significant. An important observation is that HMM systems cluster the vowel space more densely around their within-vowel means. These results are supported by the visual, correlational and statistical analysis.

3.3.2 Analysis of obstruent consonants

As mentioned in the previous section, the BC-2013 provides a variety of synthetic speech systems, which differ both in TTS technique and quality. To achieve this comparative analysis, a grouping strategy between systems was created. The explanation for each of the schemes is described below, and a concise description is displayed in Table 3.6.

Grouping strategy:-

Systems were first divided into 4 groups: *R1, R2, R3* and *R4*. *R* denotes "rank", which was decided simply by the obtained naturalness MOS for a given system. Systems that received MOS in the same interval, i.e, shared the system quality attribute, were assigned the same rank. Then, these groups were further subdivided, to explore technique-specific insights. So, all systems of the same rank and TTS technique were grouped together. Therefore, the resultant groups were: **Hybrid-R1**, **HMM-R2**, **UnS-R2**, **HMM-R3**, **UnS-R3** and **HMM-R4**, where UnS means Unit Selection. This strategy allowed us to compare high-rated systems

Rank	Group	System	Description
R1	Hybrid-R1	M K	Hybrid systems with MOS 3-4
R2	HMM-R2	I C	HMM systems with MOS 2-3
	UnS-R2	L N	UnS systems with MOS 2-3
R3	HMM-R3	H F	HMM systems with MOS 1-2
	UnS-R3	B	UnS systems with MOS 1-2
R4	HMM-R4	P	HMM systems with MOS 1

Table 3.6: Grouping strategy. Rank(R) of the system is decided by MOS for naturalness. The groups correspond to the intersection of the rank and the TTS technique (Hybrid, HMM, Unit Selection (UnS).

with low-rated systems from the same technique. HMM-R4 received poor ratings, and has not been discussed.

3.3.2.1 How does quality differ within the same TTS technique?

The purpose of this experiment is to explore quality differences between groups of the same TTS technique. The groups under comparison are HMM-R2 vs HMM-R3, and UnS-R2 vs UnS-R3. Features which showed the most statistically significant differences between groups have been identified. Comparative influences of groups on such features is presented in the subsequent sections. Boxplots Figure 3.7 and Figure 3.8 that describe the group differences have been associated with each experiment respectively.

HMM-R2 versus HMM-R3

The most informative features for observing quality differences between HMM-R2 and HMM-R3 were RMS amplitude, peak amplitude and spectral tilt.

On the basis of **RMS Amplitude**, we see differences between HMM-R2 and HMM-R3 across each manner of articulation. In affricates and fricatives, the HMM-R3 systems were observed to lower the RMS Amplitude. HMM-R2, on the other hand, did not differ significantly from the human voice in any manner of articulation. RMS Amplitude dropped in affricates by 1.8 dB, and in fricatives by 1.5 dB, with strongly significant effects (p-val $<$ 0.001). In stops, HMM-R3 systems were found to increase the amplitude by 0.51 dB, with a moderately significant effect (p-val $<$ 0.05). Therefore, through these results we can conclude that poor-quality HMM-R3 systems show lower amplitude in affricates and fricatives, and marginally higher amplitude compared to the human voice. In each case, HMM-R2 was not found significantly different from the human voice.

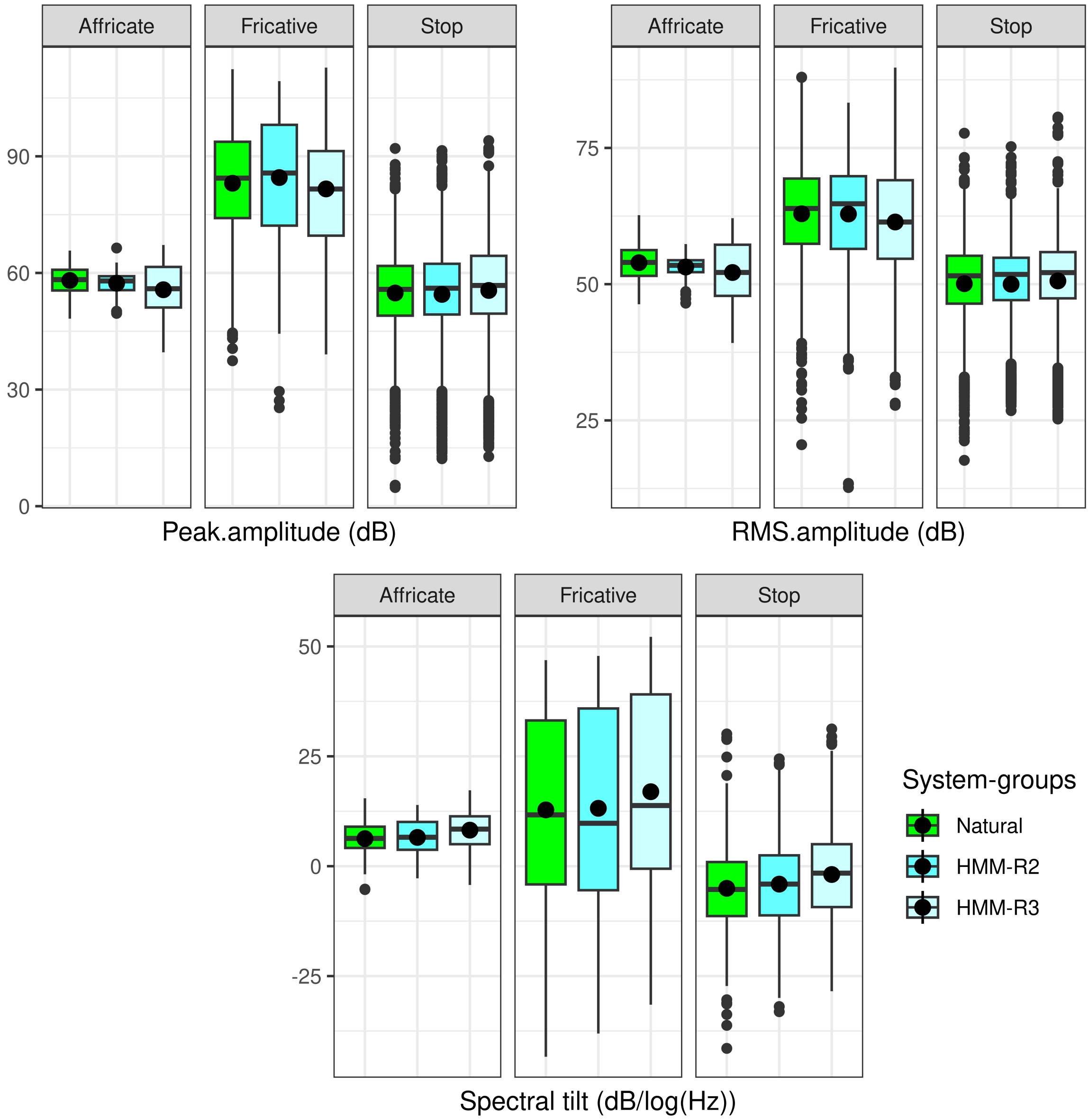


Figure 3.7: Featural comparison between system-groups HMM-R2 vs HMM-R3 arranged across manners of articulation. Median and mean values are represented with middle bar and dot, respectively.

The second feature under consideration is the **peak amplitude**. Similarly as above, HMM-R3 systems are found to lower the peak amplitude in the context of affricates and in fricatives. The peak amplitude dropped in affricates by 2.4 dB, and in fricatives by 1.4 dB, with significant effects (p-val < 0.01). HMM-R2 systems, on the other hand, do not differ from the human voice in affricates. On the contrary, they are seen to increase the amplitude for fricatives. The behaviour of the two groups was not different in stops. Therefore, we can learn that fricatives in HMM-R2 systems exhibit louder maxima of amplitude, and HMM-R3 have softer peak amplitudes in affricates and fricatives alike.

The third feature considered important is the **spectral tilt**. In all the manners of articulation, low-quality HMM-R3 systems increase the spectral tilt with strongly significant effects. The magnitude of this increase is 1.93 dB in affricates, 4.14 dB in fricatives, and 3.14 dB in stops (p-val < 0.001). In affricates and fricatives, HMM-R2 systems do not differ significantly from the human voice. But in stops, HMM-R2 also increase the spectral tilt. However, groups can still be separable within this context, because the magnitude of this increase is much lesser (0.95 dB) than in HMM-R3. Therefore, we observe that fricatives and affricates have steeper slopes in low-quality HMM systems across all manners of articulation. But in the context of stops, HMM-R2 also contribute to this effect.

UnS-R2 versus UnS-R3

The most important features for comparison between UnS groups are consonant duration, noise duration and spectral tilt.

Both UnS-R2 and UnS-R3 systems shorten the **consonant duration** in the context of fricatives and stops, while affricates do not show differences in groups for consonant duration. However, the shortening in high-quality UnS-R2 systems is seen with a stronger effect (p-val < 0.001), compared to UnS-R3 systems. In UnS-R2, fricatives are shortened by 7.5 ms and stops by 5.8 ms. In UnS-R3, on the other hand, fricatives and stops are shortened by 4.4 ms and 2.6 ms, respectively (p-val < 0.01). Therefore, we observe here that high-quality UnS-R2 systems shorten fricatives and stops more than low-quality UnS-R3.

The second feature considered important for UnS quality comparison is **noise duration**. Similar to observations for noise duration, a decrease of noise duration is found in both UnS-R2 and UnS-R3 groups for all manners of articulation. However, there are two differences. Firstly, stops show comparable decrease of noise duration between UnS-R2 and UnS-R3, and therefore are not deemed a reliable context for group differentiation. Secondly, although both fricatives and affricates have different influences of groups, they do so in different directions. UnS-R2 systems reduce the duration of fricatives with stronger significance, but affricates are shortened in UnS-R3 more strongly. Fricatives in UnS-R2 are shortened by 7.5 ms (p-val $<$ 0.001), compared to 4.4 ms in UnS-R3 (p-val < 0.01). On the other hand, affricates are shorter by 7.4 ms in UnS-R2 (p-val < 0.05), and 9.8 ms (p-val < 0.01) in UnS-R3. So here, we can learn that noise duration is reduced in both UnS-R2 and UnS-R3 groups, across all manners of articulation. group differences can be seen within fricatives and affricates. But the direction of influence is not consistent across manners.

The third feature under consideration is the **spectral tilt**. Here we see, that UnS systems on the whole lower the spectral tilt, instead of the increasing effect found in HMM systems. While the effect of lowering is strong and significant in all manners of articulation alike (p-val < 0.001), affricates and fricatives show greater separation between UnS-R2 and UnS-R3. In affricates, UnS-R2 decrease the tilt by 3.3 dB, and UnS-R3 by 7.3 dB. Similarly for fricatives,

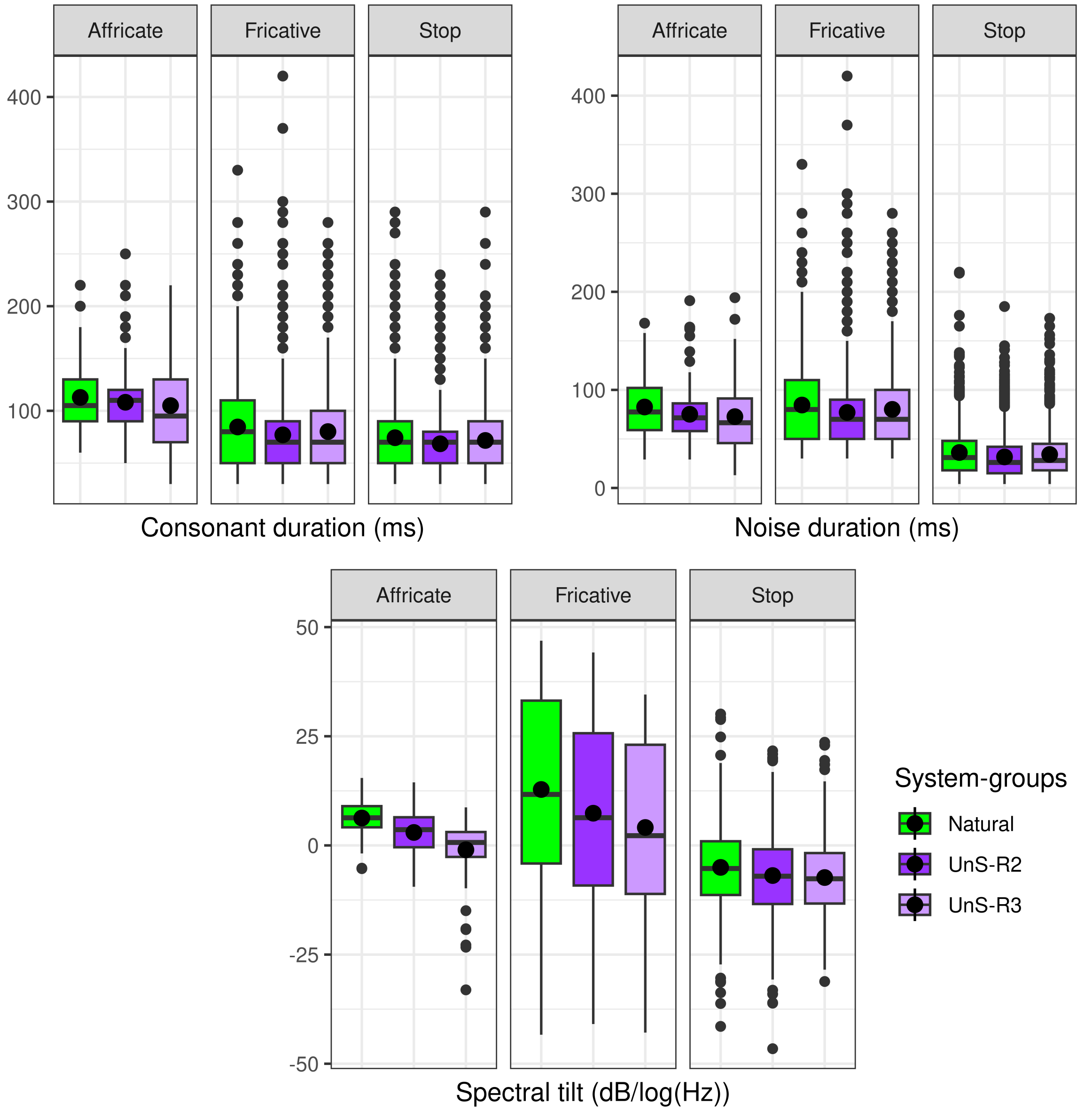


Figure 3.8: Featural comparison between system-groups UnS-R2 vs UnS-R3 arranged across manners of articulation. Median and mean values are represented with middle bar and dot, respectively.

UnS-R2 decrease the tilt by 5.43 dB, and UnS-R3 by 8.7 dB. Stops, on the other hand, show comparable lowering in both UnS-R2 and UnS-R3 groups. Therefore, this result indicates that low-quality UnS-R3 systems flatten the spectral tilt more than UnS-R2 system, especially for fricatives and affricates.

3.3.2.2 How do individual systems differ within the same quality and technique?

The purpose of this experiment is to explore individual differences between systems of the same group. Comparison will be made under Hybrid-R1 between M and K, under HMM-R2 between I and C, and under UnS-R2 between L and N. Boxplots Figure 3.9 and Figure 3.10 that describe the system differences have been associated with each experiment respectively.

System M versus System K (Hybrid-R1)

It is important to note that although M and K are in the same group, with obtained MOS of 3.9 and 3.4 respectively, that difference was statistically significant in the BC-2013 evaluations. The three most important features identified for systemic differences are RMS amplitude, peak frequency and spectral tilt.

Regarding **RMS Amplitude**, in the context of affricates, M was found to lower the RMS Amplitude by 1.7 dB (p-val < 0.001), but K was not found to be significantly different from the human voice. However, this trend completely reversed in the context of fricatives and stops. K was observed to influence a strongly significant increase the amplitude of 1.72 dB (p-val < 0.001). But in both of these contexts, M was not found different from the human voice. Therefore, affricates are softer than the human voice in M, and fricatives and stops are louder in K. So we can see that, although each manner of articulation shows systemic differences between Hybrid systems, affricates oppose the trend exhibited by fricatives and stops.

The second feature considered reliable for systemic differences within Hybrid-R1 is **peak frequency**. K shows a statistically significant raising of peak frequency in all affricates, fricatives and stops context. In affricates, the increase is by 946.23 Hz, while in fricatives, we see an increase of 337.46 Hz. Finally in stops, although the increase is smallest, of 201.8 Hz compared to other places, the effect is still strongly significant. In no context does M differ from the human voice. Therefore, K exhibits maximum amplitude at higher frequencies, while M remains closer to natural.

Finally, K shows a statistically significant raising of **spectral tilt** in each context. The increase was of 2.0 dB in affricates, 5.4 dB in fricatives, and 3.5 dB in stops. M does not differ significantly from the human voice in fricatives and stops. However, greater separation in systems can be seen in affricates, where M shows a moderately significant lowering of the spectral tilt (p-val < 0.05). Therefore, K shows a steeper slope in the spectrum, while M does not differ significantly from the human voice.

System I versus System C (HMM-R2)

Differences between I and C were **not found** in any feature, across any manner of articulation.

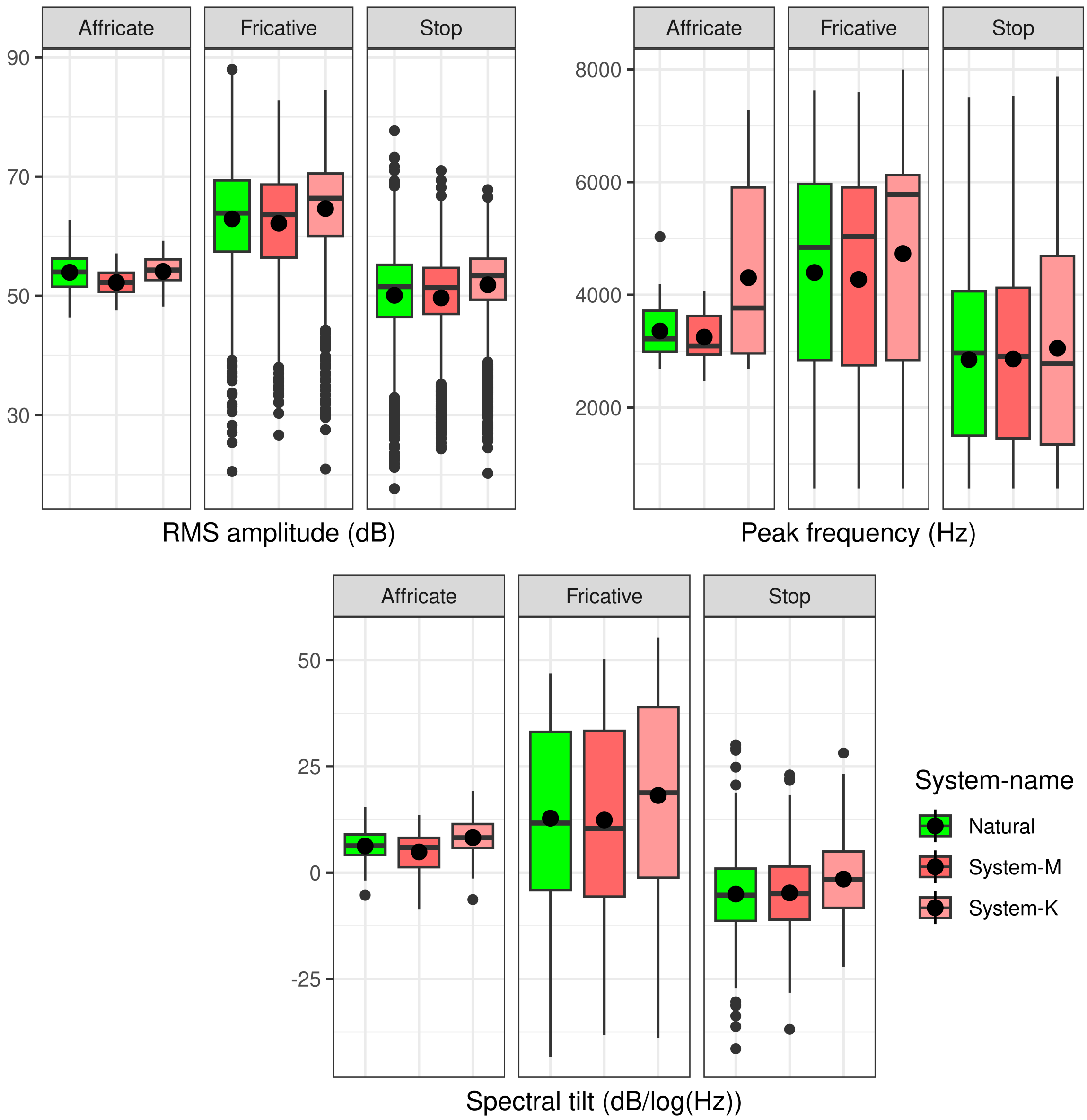


Figure 3.9: Featural comparison between individual systems (System-M vs System-K) of Hybrid-R1 arranged across manners of articulation. Median and mean values are represented with middle bar and dot, respectively.

This indicates that systems I and C have consistent patterns of influence on all the features across manners of articulation.

System L versus System N (UnS-R2)

The first feature to compare differences between L and N is **RMS Amplitude**. Differences on the basis of RMS Amplitude can be seen in all three classes of Manner - i.e., in affricates, fricatives and stops. In affricates and fricatives, N shows a strongly significant lowering of RMS Amplitude. The magnitude of this lowering is 3.0 dB and 2.9 dB in affricates and fricatives respectively (p-val $<$ 0.001). L, on the other hand, does not differ significantly from

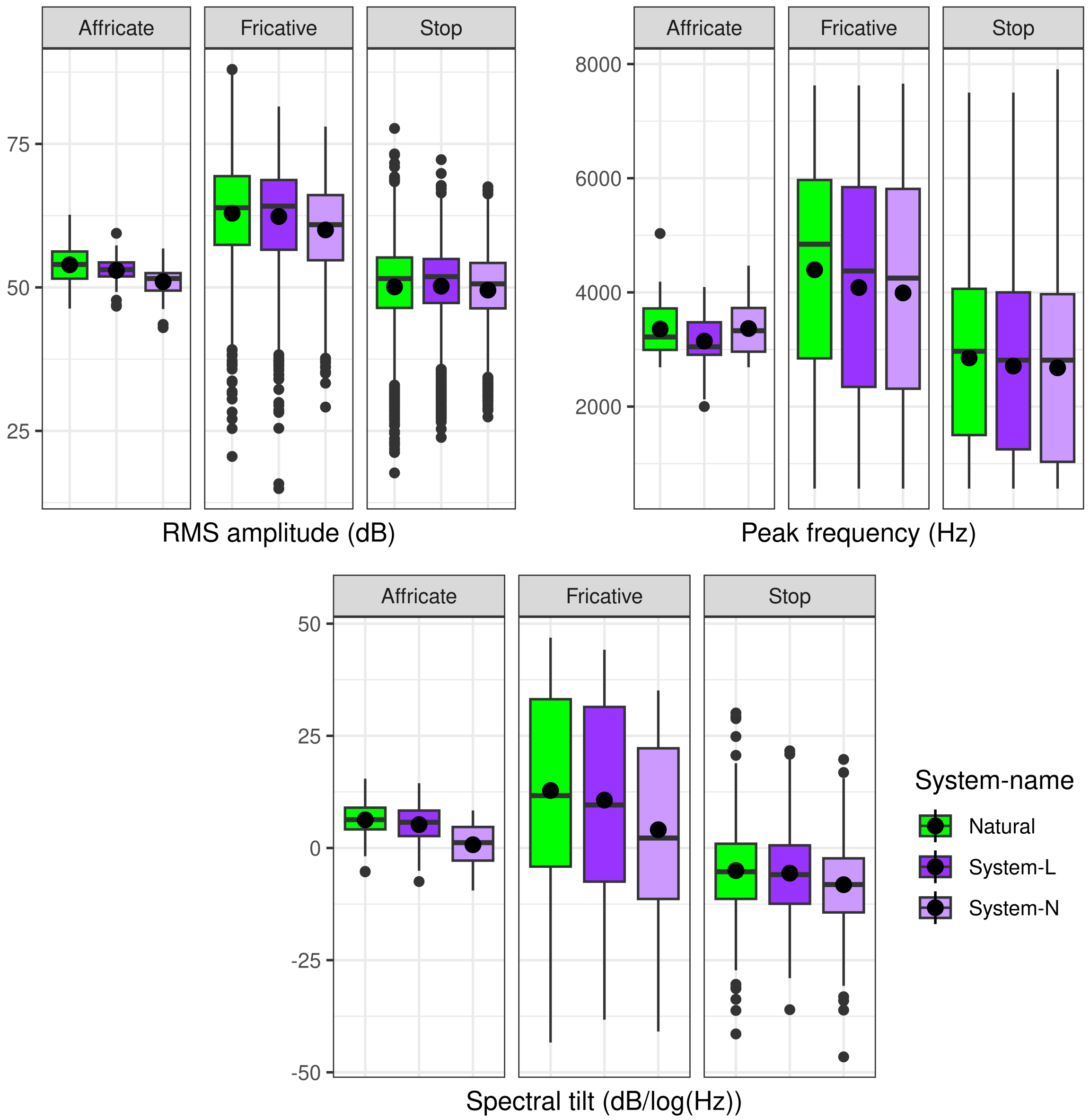


Figure 3.10: Featural comparison between individual systems (System-L vs System-N) of UnS-R2 arranged across manners of articulation. Median and mean values are represented with middle bar and dot, respectively.

the human voice. Among stops, the difference is less distinct, because N brings about only a modest lowering of 0.56 dB (p-val $<$ 0.05).

The second feature under consideration is **peak frequency**. Systemic differences can be seen predominantly in affricates, and modestly in Stops. In affricates, L shows a moderately significant lowering of 211.86 Hz (p-val $<$ 0.05), while N does not differ much from the human voice. Among stops, although the systems differ individually, the pattern of affricates is not replicated. Here, both L and N show a lowering of the frequency. The effect although, is stronger in N, with a lowering of 173.14 Hz (p-val $<$ 0.001), compared to L which lowers by 142.76 Hz (p-val $<$ 0.01).

Finally, differences based on **spectral tilt** can be seen in all three classes of Manner. In affricates and stops, N shows a strongly significant lowering of 5.55 dB (p-val $<$ 0.001) and 3.15 db (p-val $<$ 0.001) respectively, and L does not differ from the human voice. In fricatives, the difference between systems is less clearer, because both N and L show lowering. However, a greater magnitude of lowering can be observed in N, of 8.7 dB with a strongly significant effect.

Feature	**Manner**	**Differences in R2 vs R3**		**Individual differences**		
		HMM	Unit-S	Hybrid	HMM	UnS
RMS Amplitude	Affricate	R3 lower by -1.8 dB		M lowers by -1.7 dB		N lowers by -3.0 dB
	Fricative	R3 lower by -1.5 dB		K raises by +1.7 dB		N lowers by -2.9 dB
	Stop	R3 raise by +0.51 dB		K raises by +1.8 dB		N lowers by -0.56 dB
Peak Amplitude	Affricate	R3 lower by -2.4 dB				
	Fricative	R2 raise by +1.5 dB R3 lower by -1.4 dB				
	Stop					
Peak Frequency	Affricate			K raises by 946.2 Hz		L lowers by 211.9 Hz
	Fricative			K raises by 337.5 Hz		
	Stop			K raises by 201.8 Hz		L lowers by 142.7 Hz N lowers by 173.1 Hz
Consonant Dur	Affricate					
	Fricative		R2 lower by -7.5 ms R3 lower by -4.4 ms			
	Stop		R2 lower by -5.8 ms R3 lower by -2.6 ms			
Noise Dur	Affricate		R2 lower by -7.4 ms R3 lower by -9.8 ms			
	Fricative		R2 lower by -7.5 ms R3 lower by -4.4 ms			
	Stop					
Spectral Tilt	Affricate	R3 raise by +1.9 dB	R2 lower by -3.3 dB R3 lower by -7.3 dB	K raises by +2.0 dB		N lowers by -5.5 dB
	Fricative	R3 raise by +4.1 dB	R2 lower by -5.4 dB R3 lower by -8.7 dB	K raises by +5.4 dB		L lowers by -2.1 dB N lowers by -8.7 dB
	Stop	R2 raise by +0.95 dB R3 raise by +3.1 dB	R2 lower by -1.9 dB R3 lower by -2.33 dB	K raises by +3.5 dB		N lowers by -3.15 dB

Table 3.7: Feature-wise summarization of results of the linear regression model both for differences between system groups, and individual differences between systems. Each cell lists significant differences from the human voice.

Obstruent consonants: summary of results

In the obstruent consonant analysis, 10 systems from BC-2013 were grouped on the basis of their quality and TTS technique. A linear regression analysis was conducted to establish a relationship between system groups and acoustic measurements, with the human voice as reference. Features like spectral tilt, RMS amplitude were more clearly indicative of quality differences between HMM systems. Particularly, a higher spectral tilt was found associated with poor quality systems. Durational cues were more important for unit-selection groups.

3.4 Discussion and conclusion

In this chapter, we present the Dive into Divisions approach, which is an automatic and lightweight tool for evaluating TTS synthesizers. Although vowel and fricative analysis is completely automatic, adding sub-phonemic boundaries for stops and affricates has required a minimal manual supervision. Through the experiments conducted in this chapter, we provide a comparative analysis of TTS systems from the BC-2013, using contrastive features extracted from vowels and obstruent consonants. We conducted a vowel-space analysis for the steady-state portion of the vowel, and found that Hybrid and Unit-Selection systems resemble the vowel-space of natural speech more. HMM systems cluster the instances of the vowels tighter, closer to their means. A potential explanation is that HMM synthesizers generate parameters of speech using the maximum-likelihood estimation (Tokuda et al., 2000). This is an averaging step, which loses spectral detail which are irrecoverable during synthesis. In fact, techniques like re-introducing variance in the parameters has been suggested to maintain naturalness (King, 2010). In human speech, tighter clusters has been observed in clear, hyper-enunciated speech (Ménard et al., 2016; Chen et al., 2010). HMM synthesizers have been repeatedly shown to maximize intelligibility (King, 2014). These diagnostic patterns are not visible through a traditional MOS based analysis. However, studying their features provides information on their TTS technique.

Several different phenomena of quality, family and individual system differences can be observed on the basis of **spectral tilt.** In general, better-rated systems were found to be associated with spectral tilts similar to the natural voice. This is consistent with previous findings on spectral tilt contributing to improved intelligibility (Lu and Cooke, 2009). In HMM-R3 systems, spectral tilt increases from the human voice. However, in low-quality UnS-R3 systems, it is seen to decrease more steeply. Therefore, spectral tilt exhibits quality-specific differences, but the influence is dependent on the TTS technique. In terms of perceived speech quality, this indicates a preference for preserving the spectral tilt, and that deviation in either direction compromises quality. Although there is little agreement on the relationship between naturalness and intelligibility, we find that spectral tilt appears to differentiate system-groups based on naturalness as well.

Another important result can be seen is that UnS systems show differences based on quality in *durational* cues, while HMM systems on the other hand, impact spectral features more. As discussed before, statistical averaging practised in HMM systems, compromises the necessary variation required to retain spectral features. From these results, we can also speculate that the cost function of the unit selection systems favours shorter units over longer ones.

Avoiding the use of expensive behavioural equipment, we have been able to connect the domains of phonetics and speech technology. We have shown that the use of phonetic measurements is useful for a variety of comparison tasks, and the results are meaningful from a speech production and perception standpoint. A complete description of segmental properties of parallel synthetic speech can give speech synthesis researchers immediate feedback about the expectation of naturalness in their systems. These studies can precede subjective evaluation tests, by informing speech technologists about signal distortion at a segment and co-articulation level. From an acoustic-phonetic point of view, these studies allow us to understand phonemic properties that remain intact in the signal, despite a loss in naturalness. Finally, a genuine test of this approach will be in analyzing modern, *neural* TTS systems, whose quality far outperforms the systems discussed here. In the next chapter, we extend our analysis to an extended version of the BC-2013 dataset. This includes neural voices built using systems such as Tacotron (Shen et al., 2018) and FastPitch (?). It is important to note that the same speaker has been used as reference throughout, and has also provided the training data for building the neural voices. While extension of this approach with multiple speakers is proposed (See Chapter 6, Section 6.2.1 and 6.2.2), the scope in this thesis is maintained to a single speaker.

4 | The obstruent consonants of WaveNet and WaveGAN synthesizers

4.1 Introduction

In the previous chapter, we compared good-quality systems from the BC-2013 with poor-quality ones, on the basis of contrastive features of vowels and obstruent consonants. We showed that a selection of vocalic features correlate globally with the obtained MOS. Features of obstruent consonants were shown to be more informative of quality within the same technique. Quality differences between individual systems were also visible through an analysis of obstruent consonants. Therefore, the primary goal of this chapter is to use contrastive features of obstruent consonants to analyze neural TTS synthesizers.

Obstruents in this chapter are further categorized on their voicing status and their positional context, in addition to the manner distinction in the previous chapter. Transitional properties such as formant movements, relative vowel amplitude are also extracted from the vowels that appear in their neighbourhood, i.e precede or follow them. The main theme of exploration is: do features of obstruent consonants and their neighbouring vowels differ between the human voice and neural TTS voices? We find that features of voiceless fricatives and stops show more distortion than those of vocalic and voiced regions. Additionally, we also discuss which features show improvement compared to older, non-neural TTS synthesizers. The dataset comes from the recently extended version of the original BC-2013 corpus (Le Maguer et al., 2022). This version contains 4 new neural TTS synthesizers: FastPitch WaveNet, Tacotron WaveNet, FastPitch WaveGAN and Tacotron WaveGAN. Their details are presented in Section 4.3.1.

The findings of this chapter have also contributed to a parallel exploration in phonetics and speech science. The *Dive into Division* approach simply extracts contrastive features for comparing TTS voices with the human voice. But are they serving their original purpose - of communicating contrast - when produced by TTS voices? In other words, is phonemic contrast in TTS voices (e.g, /p/-/b/) maintained through the same features in human and machine voices? A complete methodological description is beyond the scope of this chapter, but the relevant findings have been described in Section 4.5.

4.2 Obstruents and their importance

We have already come across a brief introduction to obstruent consonants in Section 3.3.2. This section provides a deeper discussion to English obstruents. We will also discuss why obstruents should be given special attention in TTS, because of their role in speech perception.

4.2.1 Descriptive parameters of consonants

Phonemic differences between consonants are categorized on the basis of 3 parameters: their voicing status, place of articulation and manner of articulation. The voicing status refers to the periodic vibration of the vocal folds during their production. When vocal folds vibrate during the production of a phoneme, the resultant sound is *voiced* and a repetitive or periodic pattern can be seen in its acoustic signature. Conversely, when the vocal folds are held apart, the phoneme reflects aperiodic, noise-like characteristics in its acoustics. Voicing is contrastive in English. This means that when all the other parameters are held constant, being voiced or voiceless can change the meaning of the word. For example, in the words “seal" and “zeal", the only difference is the vibration of the vocal folds during the production of the /z/. Next, the place of articulation refers to the location of the tongue as it makes a constriction in the oral cavity. Constrictions can be anterior, i.e. made with tongue tip and blade, or posterior, i.e. made with the dorsum (or back) of the tongue rising towards the velum. There are 7 places of articulation in English that identify different consonants. For example, the difference between “sell" and “shell" is a place contrast. Finally, manner of articulation refers to the degree of constriction made as air is passing through the vocal tract. In other words, manners decide *how* the air escapes the oral tract. The broadest distinction is between ***obstruents*** and ***sonorants***. Obstruent consonants, as name suggests, obstruct the flow of air. Sonorant consonants on the other hand, allow a continuous flow of air through parallel resonating chambers such as the nasal cavity or the side chambers on the side of the tongue. Sonorant consonants are always voiced, and can be further categorized as nasals and approximants. It is important to note that voicing distinctions can be made only within obstruent consonants, at least in English. In terms of manner, further categorizations can be made in obstruents. These will be described next, and visualized in Figure 4.1.

4.2.2 Manners of articulation in obstruent consonants

- **Stops:**- Stop consonants are characterized by a complete closure, followed by a burst or “plosion" during their release. Bilabial stops are produced by forming a complete closure at the lips. Alveolars /t/ and /d/ are coronals and are produced apically, i.e. using the tongue tip. Finally, velars /k/ and /g are produced by raising the tongue dorsum is towards the velum. In English, voiceless stops also exhibit allophonic variation, such that the burst is followed by aspiration in the word-initial position. Therefore, in the word-initial position, voicing distinctions can be made using the length of this aspiration.
- **Fricatives:**- Fricatives are characterised by forcing the airstream through a narrow constriction in the vocal tract. This causes the air to escape with a high volume velocity, causing audible “frication". English fricatives broadly appear in the labiodental /f, v/, dental /θ, ð/, alveolar /s, z/ and postalveolar /ʃ, ʒ/ places of articulation. The alveolar and postalveolar fricatives also form a special class called “sibilants". In these consonants, the airstream hits the upper teeth ridge causing a *hiss* like sound.
- **Affricates:**- Affricates combine the manners of stops and fricatives such that although the air is occluded completely, its release is conducted through frication. These appear predominantly in the postalveolar place of articulation /ʧ, ʤ/.

4.2.3 Perceptual relevance of obstruent consonants

Obstruent consonants have been documented (Lindblom and Maddieson, 1988) to account for two-thirds to three-quarters of the consonantal population in cross-linguistic phoneme inventories. In the BC-2013 dataset, obstruents cover 64.01% of the consonantal population. This suggests that a large mass of acoustic cues in the utterance are comprised of obstruent consonants, which can be used by listeners while making perceptual judgments about the perceived attributes of synthetic speech.

Categorized along three manners and two voicing conditions, English obstruents also offer a set of distinctive phonetic attributes, which cannot be studied in other phonological classes. For example, effects of voicelessness may only be located within obstruents. Additionally, vowels in their neighbourhood are also influenced by their properties, and carry cues to their identification such as formant transitions (Delattre et al., 1954), F0 perturbations (Kirby and Ladd, 2016), and amplitude and duration changes (Lehiste and Peterson, 1959b; Gracco, 1994) as a function of manner and voicing.

Affricates and fricatives are known for their articulatory complexity, and their misarticulation is apparent in dysarthric speech (Kim et al., 2010) compromising intelligibility. From a speech perception perspective, obstruents are perceived less reliably in noise (Li and Loizou, 2010),

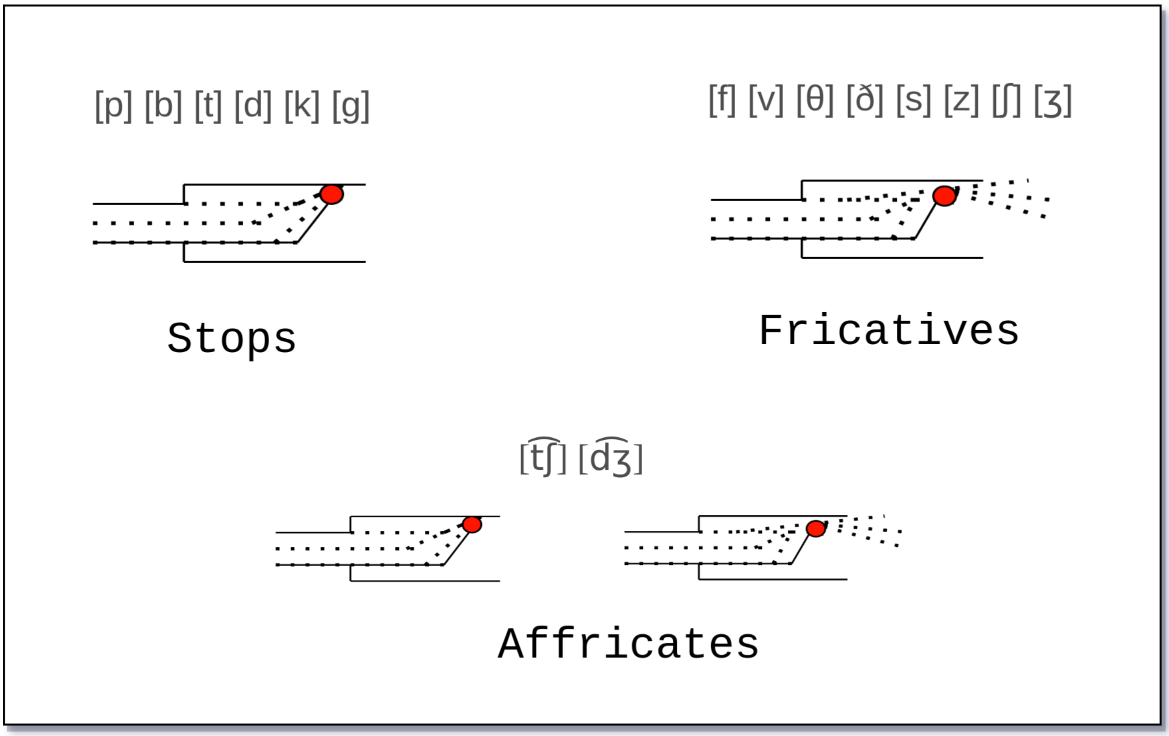


Figure 4.1: Production characteristics of obstruent consonants, categorized across the three manners of articulation (affricate, fricative and stop) found in English. Manners of articulation describe the mechanism through which air escapes the oral cavity. The red mark shows the type of alveolar tongue constriction in each manner: in stops and affricates we can see a complete closure that blocks the air, while in fricatives a narrow opening is sustained through production.

and enhancing their target cues has resulted in improved recognition of speech (Li and Allen, 2011). These studies highlight the perceptual contribution of obstruents, and the critical role their precise production plays in the perception of speech.

A targeted discussion on obstruents within TTS has been absent in the literature so far. Subjective evaluations attest that WaveNet voices sound much more natural than non-neural synthesizers (Van den Oord et al., 2016). Hence in this chapter, we hypothesize that production characteristics of obstruents in neural TTS must come closer to human speech. In other words, improvements in neural TTS may be connected to improvements in the segmental characteristics of obstruents.

4.3 Experimental setup

4.3.1 Description of the dataset

The original BC-2013 is described in Section 3.2.1, and was used for analysis in the previous chapter. The BC-2013 is an excellent resource for comparing TTS techniques on a parallel, single-speaker dataset. Additionally, the accompanying 300-hour training dataset also makes it suitable for developing modern, neural synthesizers. Therefore, it was recently extended (Le Maguer et al., 2022) to include 4 new voices: Fastpitch WaveGAN (Q), Tacotron WaveGAN (R), Fastpitch WaveNet (Y) and Tacotron WaveNet (Z). The hybrid system K, the HMM system C and the unit-selection system N were chosen as representative non-neural systems from the original challenge. A MOS based subjective evaluation showed that the neural voices scored higher on subjective naturalness compared to each of the older TTS techniques. A pairwise Wilcoxon signed rank test was used to determine the statistical significance of these MOS scores. Data was collected over 59 participants (28 female, 31 male). The new MOS scores on naturalness are summarized in the table below:-

System	Human	Z	Y	R	Q	K	N	C
MOS	4.37	3.91	3.31	3.42	3.12	2.56	1.92	1.78

Table 4.1: Mean Opinion scores on subjective naturalness as reported by (Le Maguer et al., 2022). System K, N and C are hybrid, HMM and Unit-selection systems from the original BC-2013, re-evaluated in comparison with the neural systems Y, Z (WaveNet) and Q, R (WaveGAN).

Autoregressive systems are those systems that generate the next sample based on input from the previous samples in a sequence. The autoregressive acoustic model in the extended BC-2013 is Tacotron, and the vocoder is WaveNet. **Non-autoregressive** systems generate samples of a sequence in parallel. The non-autoregressive acoustic model in the extended

	Human	Z	R	Y	Q	K	N	C
Human	NA							
Z	x	NA						
R	x	x	NA					
Y	x	x		NA				
Q	x	x			NA			
K	x	x	x	x	x	NA		
N	x	x	x	x	x	x	NA	
C	x	x	x	x	x	x		NA

Table 4.2: Significance results of a pairwise Wilcoxon signed rank test with Bonferroni correction for our listening test. Each cell marked with **x** indicates that the two systems are considered different with a p-value < 0.01.

BC-2013 is FastPitch, and the vocoder is WaveGAN. This makes the extended BC-2013 extremely advantageous, because it provides a 4-way cross-combination of autoregressive and non-autoregressive TTS generation techniques.

4.3.1.1 The autoregressive vocoder - WaveNet

WaveNet (Van den Oord et al., 2016) is an autoregressive generative neural architecture which uses dilated causal convolutions for generating high-quality audio. The model is composed of stacked convolution layers with no pooling steps, and a softmax layer to predict the output. Each sample is predicted as a categorical distribution based on the conditional probabilities of all the previous samples in the waveform. The use of causal convolutions ensures that the predicted sample depends only on the previous samples. To reduce computational cost and time complexity, dilated convolutions are used. This means that units of the input are incrementally skipped while stacking the convolutional layers. Categorical distribution, particularly softmax, is recommended for modelling conditional probabilities, despite the continuous nature of input in speech. This is because greater flexibility was observed, compared to mixture networks where data had to fit certain shape assumptions. For activation functions, gated units were preferred to linear ones based on better performance. A useful parameter, the conditional h is also added to the conditional probability distribution in addition to the previous timesteps. This enables WaveNet to adapt to multiple speakers, or switch between speech tasks (voice conversion, text-to-speech). The local conditioning feature is specially important for TTS, as linguistic features (log F_0, phone duration) were extracted from the text.

While WaveNet produced high-quality audio, its speed was too slow for real-time applications. Improvements in speed and scalability are proposed through Parallel WaveNet (Oord et al., 2018) and ClariNet (Ping et al., 2018). In particular, parallel WaveNet, presents a dual approach where the sampling procedure of IAF networks and the parallelizable, convolutional training procedure of the original WaveNet are combined in a teacher-student framework. Given a sample x_t, the IAF can infer the output at previous timesteps, therefore generate all samples in parallel. This forms the student part of the network, which learns from the original WaveNet, the teacher.

4.3.1.2 The non-autoregressive vocoder - WaveGAN

The non-autoregressive vocoder used in the extended BC-2013 is the WaveGAN (Yamamoto et al., 2020). The autoregressive neural generation techniques showed promise in increasing the processing time and complexity. However, the inference speed was still slow and computationally expensive. Additionally, a trial-and-error method was used to optimize the density distillation process. Therefore, Parallel WaveGAN (Yamamoto et al., 2020) was introduced to sidestep the traditional teacher-student training methodology, and also overcome the distillation process. WaveGAN also preserves the perceived TTS quality by providing a straightforward estimation of the waveform, instead of linear prediction and conversion approach. Instead, a "joint optimization" method is proposed. This involves a linear combination of adversarial loss, and the multi-resolution STFT loss functions. The adversarial loss represents the loss function of the generator part of the adversarial network. The generator is designed to produce samples which deceive the discriminator, and also capture the underlying distribution of speech waveforms in the process.

Having a multi-resolution loss means combining individual loss from multiple analysis parameters, such as FFT size, window size and frame shift. This allows for greater flexibility, as representation from multiple parameters reduces generator overfitting. While WaveGAN also uses dilated convolutions, their relationship is not causal.

4.3.1.3 The autoregressive acoustic model - Tacotron

The main autoregressive acoustic model architecture used in the extended BC-2013 is the Tacotron (Wang et al., 2017). Tacotron provides an end-to-end generative process for synthesizing spectrograms directly from text, which can then be converted to waveforms. It employs a sequence-to-sequence encoder model with an attention-based decoder. The purpose of the encoder module is to extract representations from character sequences. For this, the character input is passed through convolution filters, which model the target phoneme and its context. The outputs are max pooled, incremented with the original character sequence and passed into a highway network for extracting high-level features. A gated unit further extracts features from both the preceding and succeeding contexts. The task of the decoder is to convert features into spectrograms. First, a context vector and the output of a recurrent layer is concatenated to form the input to the decoder. Although raw spectrograms can be predicted, an 80-band mel-spectrogram is the chosen target so that proper alignment can be learnt between the speech signal and text. Additionally, multiple non-overlapping frames are predicted together, utilizing the relationship between neighbouring phonemes and also speeding up the prediction process. Finally a post-processing net converts the output of a decoder to a representation that can be passed to a vocoder.

4.3.1.4 The non-autoregressive acoustic model - FastPitch

The main architecture used in the extended BC-2013 is the FastPitch (Łańcucki, 2021), which predominantly builds on the FastSpeech Ren et al. (2020) architectures. The model employs a stacked, length regulated feed-forward transformer structure to predict mel-spectrograms from phoneme sequences. This structure comprises of self-attention and a 1-D convolutional network. The self-attention network is built using a multi-head attention layer which can detect cross-positional information of the phoneme sequences. Similarly, the 1-D convolutional network is designed to take advantage of the relatively closer relationship between the hidden states, when modelling phonemes and mel-spectrograms speech. An important contribution of the Feed Forward Transformer (FFTr) network is the length regulator sequence. It addresses the length mismatch in the phoneme and mel-spectrogram sequence. Specifically, the length of the mel-spectrogram sequence is longer than the phoneme sequence, because multiple mel-spectrogram sequences may collectively correspond to a single phoneme. The regulator expands the length of phoneme sequence to match that of the mel-spectrogram sequence, thus balancing the mismatch. Additionally, this allows for greater prosodic and pause insertion flexibility. Another important detail of the FastSpeech model is the duration predictor network, that predicts the number of mel-spectrograms required for one phoneme, or its duration. A separate 2-layer 1-D convolutional network is trained on a an autoregressive Transformer model, and used using TTS inference time.

Thus, the dataset for analysis comprised of these 4 voices (systems Y, Z, Q and R), 6 top-quality non-neural voices and the human analysis. Each system provided an identical set of 100 sentences, which formed the parallel data for comparison. Features were independently extracted from each voice, compared statistically against the human voice as the reference. The feature extraction and statistical analysis steps are described below.

4.3.2 Feature extraction

Contrastive properties of obstruent consonants have been studied along the durational (Jongman, 1989), amplitudinal and spectral (Chodroff and Wilson, 2014; Stevens and Blumstein, 1978) attributes. The perceptual contribution of their surrounding vowels has also been discussed (Sussman et al., 1995).

Audio files from all the systems were force-aligned using the MFA (McAuliffe et al., 2017) to create phoneme-level boundaries. Sub-phonemic boundaries for the noise duration of stops and affricates were demarcated using a rule-based temporal boundary identification procedure described in Section 3.2.3.2. Consonants were then separated into 3 positional contexts: pre-vocalic (CV), post-vocalic (VC) and consonant clusters. Consonant clusters were not analyzed for the present analysis. Vowels that appeared in the immediate neighbourhood of these consonants were also categorized into the CV and VC positional contexts. Then, the following feature set was extracted for the **analysis of vowels**:-

- **Vowel duration (V-Dur):-** The duration of the vocalic region, as returned by the MFA. In the CV position, the vowel onset is marked at the first 20% of this duration. Conversely, in the VC position, the offset is marked at the last 20% of this duration.
- **RMS amplitude (RMS-Amp):-** The root-mean-squared amplitude of the power spectrum of the vocalic region.
- *Formant values (F1-F5):-* The formant values of the first 5 formants at the onset/offset and midpoint of the vowel. These were extracted using the Burg formant-tracking algorithm in Praat (Boersma and Weenink, 2018). The Escudero optimization procedure (Escudero et al., 2009) was used to estimate the appropriate ceiling value.
- **Within-category dispersion:-** The absolute difference between formant values of individual instances of the vowel, and the mean of formants across all the instances of that vowel. Dispersion values were calculated for formants at both onset/offset *(On-Fn-disp)* and midpoint *(Mid-Fn-disp)*.
- **Relative amplitude (F_n-RA):-** The difference between amplitude of the vowel spectrum at F3, F4 and F5, and the consonant at the corresponding frequency.
- **Spectral tilt (Sp-Tilt):-** The slope of the least-squares regression line fitted after log-transforming the frequency domain of the spectrum.

For the **analysis of consonants**, the feature extraction procedure was identical to the 8 features described in (Pandey et al., 2021). The features were:- *consonant duration*; *noise duration*; *RMS amplitude*; *peak amplitude*; *peak frequency*; *dynamic amplitude*; and *spectral tilt*. In addition to these features, the present analysis also included *spectral shape* for consonantal analysis. Spectral shape has been described (Evers et al., 1998) as the difference between the spectral tilts below and above the mid-frequency region. (Tilt$<$ 2.5 kHz - Tilt$>$ 2.5 kHz). All of these features were analyzed separately in their positional contexts (CV, VC), as opposed to their global evaluation in (Pandey et al., 2021).

4.3.3 Statistical analysis

The feature set failed the Shapiro-Wilk test for normality, and also had unequal variances among the groups. Therefore, we decided to use non-parametric tests: the Kruskal-Wallis (Kruskal and Wallis, 1952) test, and the Dunn Test (Dunn, 1964).

4.3.3.1 The Kruskal-Wallis test - grouped comparisons

The Kruskal-Wallis test (Kruskal and Wallis, 1952) examines group differences within a population, especially when the data does not follow a parametric distribution. In other words, it compares categorical differences within non-normal datasets. It was proposed as an alternative for the parametric analysis-of-variance methods, and used *ranks* in the data, instead of the observed values themselves. Observations regardless of their group membership are arranged in ascending order based on their numeric value. Then, the test statistic H is computed based on the following equation:-

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2} \tag{1}$$

Here, r_{ij} is the rank of each observation j in group i, $\bar{r}_i$ is the average ranks within the group i and $\bar{r}$ is the average of all the r_{ij}s. N and g stands for number of observations and groups respectively. As stated in the original paper, "Large values of H generally mean significant results."

The H statistic computed is compared against a standard chi-square distribution with the same degrees of freedom. To reject the null hypothesis, the computed H statistic must be greater than H_c, at a chosen significance level.

$$H >= H_c \quad \text{at} \quad p - value < 0.05 \tag{2}$$

For example, the critical value H_c for a chi-squared distribution is 7.815 at 3 degrees of freedom, for a significance level of 0.05. If our computed H statistic is greater or equal to this value, then the null hypothesis can be rejected. Throughout this chapter, the values of H statistic and the degrees of freedom will be reported. Associated probability values (p-val) will also be reported to indicate the strength of significance. It must be noted that Kruskal-Wallis is a test of **grouped** data. For example, if observations from the human voice, HMM synthesizers and neural synthesizers are put together, Kruskal-Wallis can identify that there are statistically significant differences between the three techniques. However, how each technique is deviates from the human voice is a task for a post-hoc analysis using the Dunn's test. This is described next.

4.3.3.2 The Dunn's test - pairwise comparisons

The Dunn's test for pairwise comparisons was developed by the American scientist Olivia J Dunn in 1964 (Dunn, 1964). While the Kruskal-Wallis can identify the presence of significant differences among the groups in the populateion, the Dunn's test pairwise compares each of the groups, and identify those which differ from each other.

This is two-step procedure. First, the difference y_m is computed as the difference between the averaged ranks of the two groups that we want to compare. If A and B are the two groups, then y_m is described by the following equation:-

$$y_m = \bar{r}_A - \bar{r}_B \tag{3}$$

The $\bar{r}_A$ and $\bar{r}_B$ are the averaged ranks of each group and correspond to the $\bar{r}_i$ value obtained in the Kruskal-Wallis test. Second, this y_m is divided by its own standard deviation σ_m. The following equation describes the squared σ_m^2:-

$$\sigma_m^2 = \left[\frac{N(N+1)}{12} - \frac{\sum_{k=1}^{g}(t_k^3 - t_k)}{12(N-1)}\right]\left(\frac{1}{\sum_a n_A} - \frac{1}{\sum_b n_B}\right) \tag{4}$$

As before, N and g represent the total number of observations and groups respectively. Here, t_k depends on the previously computed r_{ij}, such that it represents the number of times the r_{ij} appears in the dataset. It is subtracted from its cubed form t_k^3 and aggregated over g, i.e., the total number of groups. The values n_A and n_B are the total number of observations in the groups under comparison.

The resultant standardized difference y_m/σ_m is a probability (p) value. At a chosen significance level (usually, $p < 0.05$), we can compare it with the critical range of the z-score table. If our obtained p-value falls outside this range, then the groups are significantly different.

Thus, a pairwise comparison is performed between the human voice and each TTS technique or system. This helps to determine those features where the differences between the human and TTS voices are statistically significant.

4.4 Results

In this section, we describe the results of statistical analysis from the consonantal and vocalic feature-set. First, we explore those segmental features which show improvement compared to older, non-neural TTS techniques. Then, we identify those features which deviate from the human voice in neural vocoders. While global trends are discussed, a system-specific comparison is detailed for each neural voice and the human one.

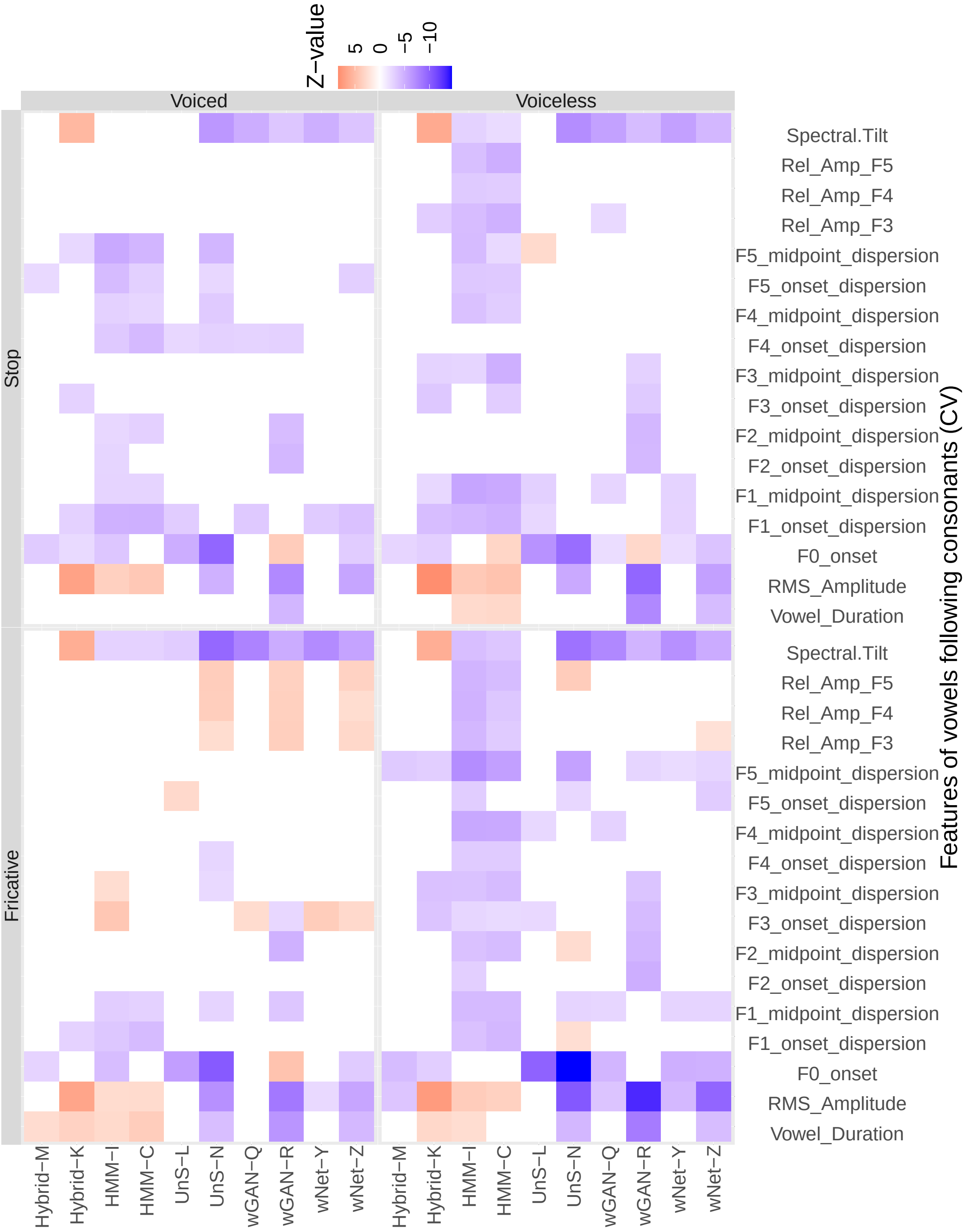


Figure 4.2: **Vowels following consonants (CV):** Deviation in acoustic-phonetic features of vowels following (CV) obstruent consonant. Deviation is calculated as statistically significant difference between the human voice and each TTS system of the BC-2013 dataset, in a pairwise Dunn's test. Dark cells indicates strongly significant differences, while white cells represent no significant difference compared to the human voice.

4.4.1.1 Visual analysis: trends in vowels

Figure 4.2 and Figure 4.3 show the features which deviate from the human voice on the basis of Dunn's test. Coloured cells show statistically significant differences, while white cells mean no difference from the human voice.

In HMM synthesizers I and C, we see a blanket lowering of within-category dispersion in lower

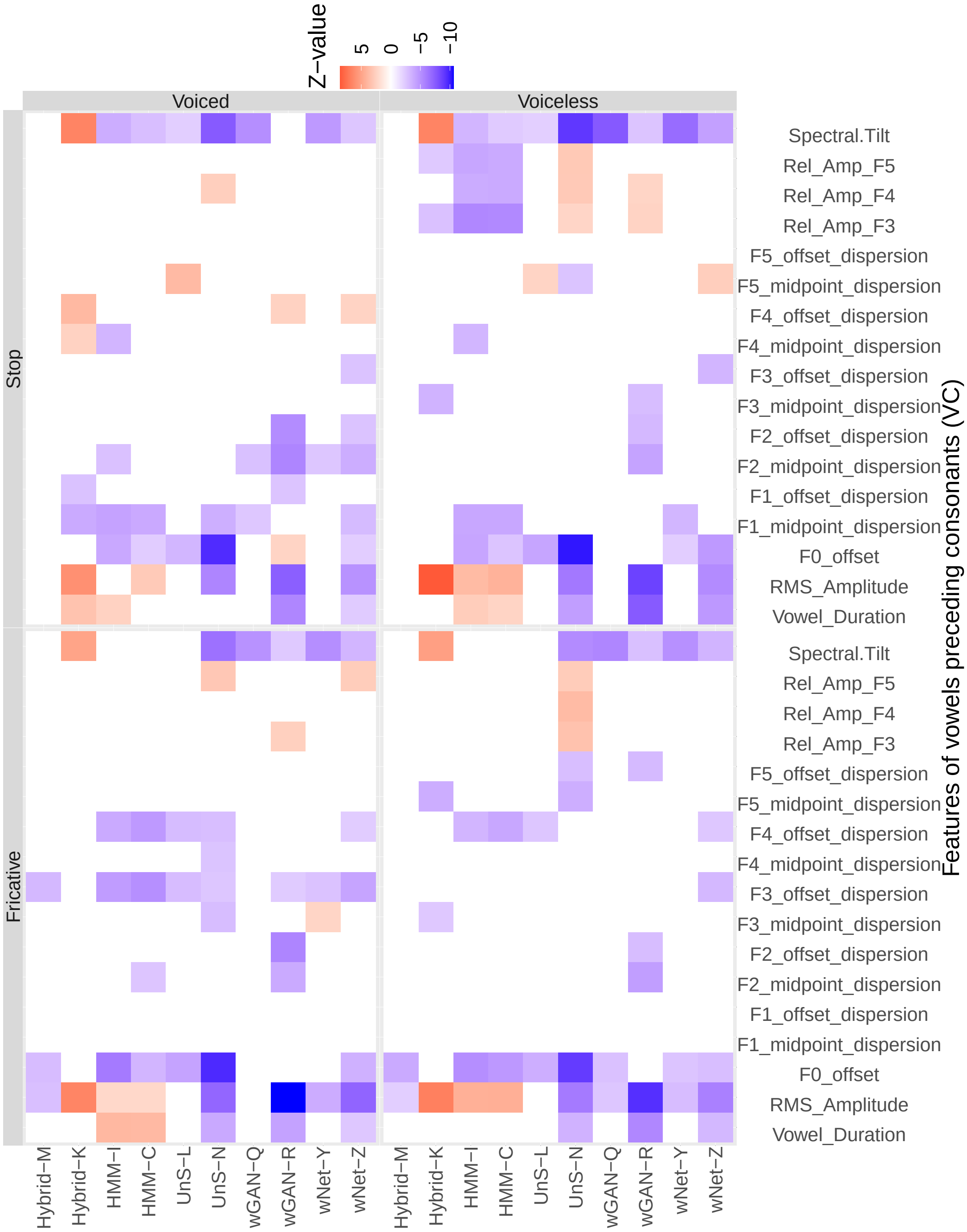


Figure 4.3: **Vowels preceding consonants (VC):** Deviation in acoustic-phonetic features of preceding (VC) an obstruent consonant. Deviation is calculated as statistically significant difference between the human voice and each TTS system of the BC-2013 dataset, in a pairwise Dunn's test. Dark cells indicates strongly significant differences, while white cells represent no significant difference compared to the human voice.

and higher formants. This is especially visible in the CV position especially for voiceless stops and fricatives. Secondly, I and C also lower the relative amplitude in vowels. In comparison, neural synthesizers Q, R, Y and Z show improvement in both these featuers, and show no statistically significant deviation from the human voice.

Compared to Unit-selection L and N, and Hybrid synthesizers M, K, also we see improvement on a select set of segmental features. Specifically, even the top-quality unit-selection synthesizer L shows a lowered F0 onset in every vocalic context. Neural systems Q and Y maintain human-like F0 onset, at least in the voiced contexts. It must be noted, that since concatenative synthesizers use units directly from human voice recordings, these voices are segmentally closer to the human voice than parametric synthesis.

In neural synthesizers, a distinct trend emerges from the lowering of spectral tilt, visible in all the vocalic contexts. Although statistically significant in every neural voice, FastPitch systems (Q, and Y) show a sharper dip compared to Tacotron. Darker cells are visible in both the CV and VC contexts. Therefore, from Figure 4.3 and Figure 4.3 we gather that although neural voices show improvement in a variety of segmental features, but some features consistently deviate from the human voice.

4.4.1.2 Vowels: neural versus non-neural TTS

Compared to HMM systems I and C, neural voices show a variety of important differences. First, as seen in the previous chapter, HMM voices lower the within-category dispersion. Positional and voicing influences can be seen, as the effects are most consistent in vowels following the voiceless obstruents (i.e, CV). Although all formant values show a significant lowering, higher formants are impacted more. Dispersion at F5 midpoint drops by a median of 159.75 Hz, and differs strongly from the human voice [$\chi^2(1) = 75.37$, p-val < 0.0001]. Similarly, at F4 it drops by 139.99 Hz. Vowels that follow voiceless fricatives show the lowest within-category dispersion values in HMM synthesizers.

In most neural voices, dispersion patterns do not show statistically significant deviation in within-category dispersion. This means that in a majority of cases, neural voices have overcome the tight, within-category clusters that were characteristic of HMM synthesizers. Some effects of clustering are seen in system R, but only in lower formants (F1-F3). The maximum reduction observed is only of 83.22 Hz at the F2 midpoint for VC vowels [p-val < 0.0001]. This shows that the magnitude of lowering is notably lesser, compared to HMM synthesizers I and C.

Neural voices also show improvement in terms of relative amplitude, when compared with HMM synthesizers I and C. Relative amplitude signifies the amplitude difference between the consonantal and vowel region. This relationship is important for proper production of sibilant fricatives, as it cues place contrast. The difference is found statistically reduced in HMM synthesizers. Vowels neighbouring voiceless stops in both positional contexts exhibit this trend, but VC shows greater median reduction of 5.24 dB [$\chi^2(1)$ = 31.07, p-val $<$ 0.0001]. In neural synthesizers, human-like patterns of relative amplitude are generally maintained. Only a a minor increase of 2.54 dB [p-val $<$ 0.05] is seen in vowels preceding voiceless stops. Other neural systems do not deviate from the human voice.

Finally, compared to unit-selection voices L and N, we can consider the case of F0 onset. Unit-selection synthesizers lower the F0 in both positional contexts. The maximum lowering of 32.38 Hz can be observed in the vowels following voiceless fricatives, with strongly significant effects [$\chi^2(1)$ = 301.58, p-val $<$ 0.0001]. Although lowering of F0 can be seen in neural system Z too [p-val < 0.0001], its magnitude 15.37 Hz is halved at the same positional and voicing counterpart. This means that unit-selection synthesizers lower the F0 at onset with much greater magnitude and statistical significance, compared to neural ones.

4.4.1.3 Vowels: neural TTS vs the human voice

Only a few vocalic features in neural TTS deviate from the human voice. However, global tendencies, and those specific to each acoustic model can be identified.

First, a statistically significant lowering of the **spectral tilt** is observed in every vocalic context, unconditioned by positional or voicing status of the adjacent vowel. Overall, spectral tilt is reduced by 2.61 dB/log(Hz), with strongly significant group effects [$\chi^2(1)$ = 205.99, p-val $<$ 0.001]. Maximum lowering of 3.02 dB/log(Hz) is seen in vowels that follow voiced fricatives, and is strongly significant [p-val $<$ 0.001]. Similar effects are observed in the voiceless condition, with a median reduction 2.71 db/log(Hz) [p-val < 0.001]. This means that, although significant cross-contextually, vowels that follow fricatives contribute maximally to the lowering of spectral tilt.

We also observe that FastPitch exhibits an even greater lowering of the tilt. Darker cells are visible in Figure 4.3 and Figure 4.3 for FastPitch systems i.e, Q and Y. Combined over manners and voicing conditions in the CV context, FastPitch systems lower the spectral tilt by 3.14 db/log(Hz), while Tacotron systems by 2.12 db/log(Hz). And specifically in vowels that follow voiceless fricatives, FastPitch show lowering by 3.47 db/log(Hz), while Tacotron systems by 2.09 db/log(Hz). In the VC context, the difference between FastPitch and Tacotron is clearest in the voiced stops condition. Here, FastPitch brings the tilt down by 2.35 db/log(Hz), while Tacotron only by 1.23 db/log(Hz). Additionally, while FastPitch exerts strongly significant effects [p-val < 0.001], Tacotron maintains only a minimal level of statistical significance [p-val < 0.05]. So, spectral tilt lowering is an informative site for both overarching tendencies in neural synthesizers, and also effects specific to the acoustic model.

Other sites for locating acoustic-model specific effects are **RMS amplitude.** Darker cells are visible for R and Z, i.e., Tacotron systems (see Figure 4.2). In stops, only Tacotron lowers the amplitude, where FastPitch is not statistically different from the human voice. Between Tacotron systems, system R maximally lowers amplitude in every context. In vowels that precede voiced stops (VC), Tacotron reduces the amplitude by 3.55 dB [p-val < 0.001]. System R on its own contributes a significant lowering of 4.02 dB [p-val < 0.001]. On the other hand, system Q and Y maintain human-like amplitude in stops. They differ minimally in fricatives, such that Q reduces amplitude by -0.96 dB, and Y by -1.11 dB. The strength of statistical significance is also modest, compared to strong effects observed in Tacotron systems.

Finally, system R shows significant reduction in the **within-category dispersion**, which is normally observed for HMM systems. This effect is the maximum in vowels that precede voiced stops (VC), where the midpoint dispersion is reduced by 83.22 Hz [p-val < 0.001]. Similarly, in the CV position, system R reduces within category variation by 62.25 Hz [p-val < 0.001]. However, higher formants are not majorly impacted, and the magnitude of this reduction is also much lesser compared to HMM synthesizers.

4.4.1.4 Summary of vowel analysis

Through this analysis, we can see that neural synthesizers resemble the human voice over a variety of vocalic features, extracted from vowels that appear in the neighbourhood of obstruent consonants. Considerable improvements can be seen compared to non-neural synthesizers, in terms of within-category dispersion, relative amplitude and F0 onset. However, features like spectral tilt and RMS amplitude still exhibit deviation from the human voice.

4.4.2 Analysis of obstruent consonants

4.4.2.1 Obstruents: visual analysis

As in the case of vowel analysis, here we will describe the major trends that we can pick up from Figure 4.4 and Figure 4.5. Then, we will support it with quantitative measures, as a result of Kruskal-Wallis and the post-hoc Dunn's test.

In neural synthesizers Figure 4.4 and Figure 4.5 clearly show major differences in the voiced and voiceless regions of obstruent consonants. Most acoustic-phonetic features of voiceless obstruents show deviation from the human voice. Particularly, system R shows deviation in more features compared to other neural synthesizers. For example, in the voiceless fricative context, every feature is impaired. Distortion in voiceless obstruents can be seen as an overarching trend, evident in both fricatives and stops, in both the pre-vocalic (CV), and post-vocalic (VC) positions, for both WaveNet and WaveGAN vocoders. It is quite apparent then, that neural synthesizers poorly reproduce the features of voicless obstruents.

In voiced obstruents, however, we can identify several improvements that neural synthesizers have made, compared to older, non-neural TTS synthesizers. For example, in voiced fricatives, we find that dynamic amplitude has been reduced compared to the human voice. In the same positional and voicing context, neural synthesizers match the human voice more closely. Similarly consider the case of unit-selection synthesizers. The duration of consonants is shortened in both unit-selection systems L and N. In voiced obstruents especially in the CV position, we can see that obstruents do not deviate.

Therefore, from Figure 4.4 and Figure 4.5 we can mark voiceless obstruents as an important site where neural synthesizers can be seen to deviate from the human voice. In voiced obstruents, improvements over non-neural TTS synthesizers can be seen over a variety of features.

4.4.2.2 Obstruents: neural versus non-neural TTS

As can be seen from Figure 4.4 and Figure 4.5, neural voices maintain the features of voiced obstruent consonants fairly well. There are several improvements in these features compared to older, non-neural TTS systems. The first feature under consideration is **dynamic amplitude**. In HMM synthesizers, I and C, we can see a statistically significant lowering of the dynamic amplitude in the pre-vocalic (CV) voiced fricative. The magnitude of this difference is 1.94 dB, with strongly significant effects [$\chi^2(1)$= 1.04, p-val < 0.001]. Since dynamic amplitude is the difference between the high frequency peak and the low frequency trough, the reduction means that the range of amplitudes in the obstruent spectrum has been restricted. Comparatively in neural voices, this range is not statistically different from the human voice.

Unit-selection voices L and N show reduced consonant duration, compared to the human voice [$\chi^2(1)$= 30.78, p-val < 0.001]. The shortening of duration can impact the contrastive perception in sibilant fricatives, as longer durations are associated with sibilance. The maximum effect is contributed by System N. System N shortens this duration by 11.99[1] ms in voiced fricatives, and 8.72 ms in voiced stops. These effects are strongly significant in voiced fricatives (p-val < 0.0001), and consistent in voiced stops (0.01). These trends are maintained in the VC position as well, with durations reduced in voiced fricatives by 7.51 ms (p-val < 0.05), and stops by 8.77 ms (p-val < 0.01). The duration of neural synthesizers also differs from the human voice. However, as Figure 4.3 and Figure 4.3 clearly shows, this is limited to voiceless conditions in both CV and VC conditions. In voiced fricatives and stops, the reduction is only by a 3.60 ms and 5.11 ms respectively. In neither case is this significant. Furthermore, comparing individual systems, we also note that the most contributive effects are localized to system R. Although

[1] average values reported here, median subject to rounding error

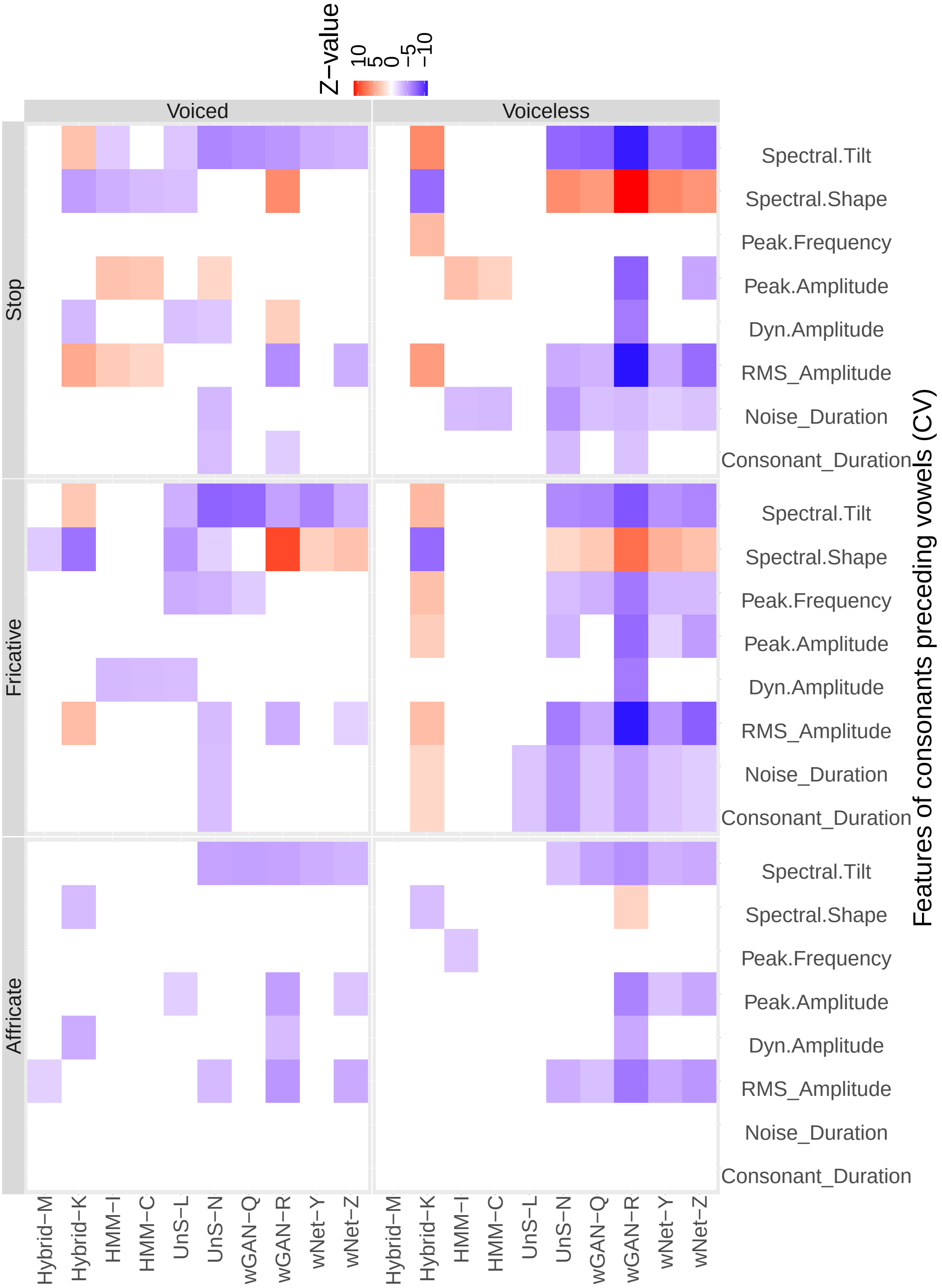


Figure 4.4: **Consonants preceding vowels (CV):** Deviation in acoustic-phonetic features of preceding obstruent consonants (CV). Deviation is calculated as statistically significant difference between the human voice and each TTS system of the BC-2013 dataset, in a pairwise Dunn's test. Dark cells indicates strongly significant differences, while white cells represent no significant difference compared to the human voice.

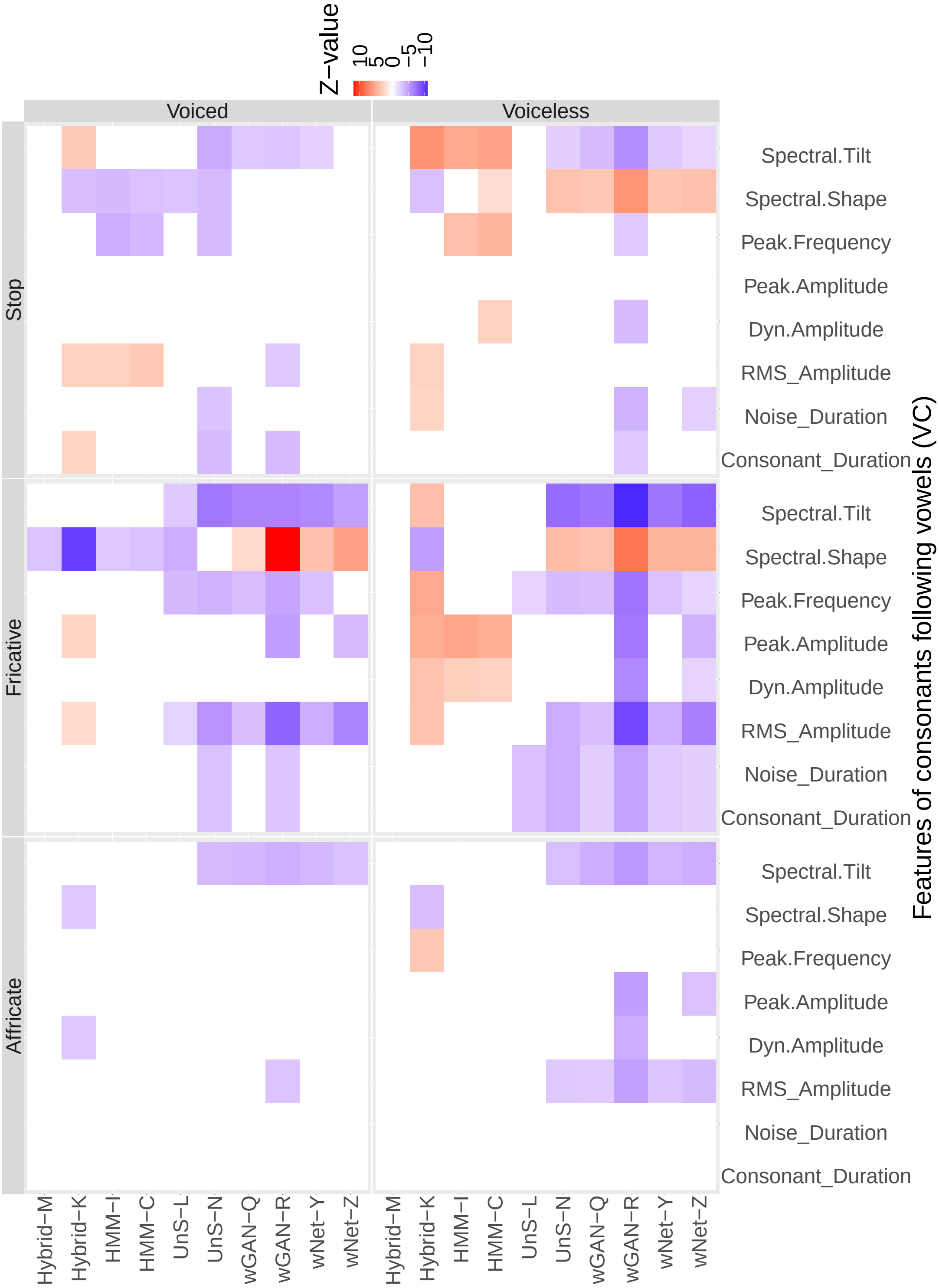


Figure 4.5: **Consonants following vowels (VC):** Deviation in acoustic-phonetic features of following consonants (VC). Deviation is calculated as statistically significant difference between the human voice and each TTS system of the BC-2013 dataset, in a pairwise Dunn's test. Dark cells indicates strongly significant differences, while white cells represent no significant difference compared to the human voice.

all neural voices exhibit relatively shorter durations, the difference is not significant in any except R.

The next feature under consideration is the RMS amplitude of voiced stops, both in the CV and VC position. As the amplitude of the burst spectrum is a useful place classification cue, its increase may impact place perception in voiced stops. Hybrid synthesizers increase the RMS amplitude [χ^2(1)= 8.45, p-val $<$ 0.01]. Similar trends are exhibited by the HMM synthesizers [χ^2(1)= 30.78, p-val $<$ 0.001]. Upon comparing individual synthesizers, we find that Hybrid system K increases the amplitude by a median of 1.92 dB, with strongly significant effects [p-val < 0.0001]. Both HMM synthesizers also show significant increases, but system I [p-val < 0.01] has a slightly stronger effect than system C [p-val < 0.05]. These values are reported for the CV context, but the trends are comparable in the VC context. Conversely, FastPitch synthesizers do not differ from the human voice in any positional context. Tacotron synthesizers do exhibit lowering, but only system R is significant in both positional contexts.

4.4.2.3 Obstruents: neural TTS vs the human voice

From Figure 4.4 and Figure 4.5 we see that **voiceless** fricatives and stops show divergence from the human voice across several features. This trend is visible in both CV and VC contexts, and is more prominent in the CV context. This indicates a broad, overall tendency of neural systems to model characteristics of voiced obstruents better than voiceless ones.

The most important featural divergence can be observed in terms of the consonantal spectral tilt. As seen in vocalic features, this extends to both positional contexts, voicing conditions and manners. This means that in neural voices, high frequency regions of the consonants are more damped than they are in the human voice. On the whole, neural voices lower the spectral tilt by 4.82 dB/log(Hz) in the post-vocalic CV context, with strong group differences [χ^2(1)= 173.7, p-val $<$ 0.0001]. Parallel trends are observed in the VC context, with a median 3.0 dB/log(Hz), statistically significant lowering [χ^2(1)= 56.78, p-val $<$ 0.0001]. In voiceless fricatives especially, we see a greatest drop of 9.53 dB/log(Hz) with strongly significant effects [p-val < 0.0001]. In the VC obstruents, this is even sharper, exhibiting a median lowering of 11.18 dB/log(Hz). The neighbourhood of voiceless fricatives was noted as an important site for vocal spectral tilt reduction, as seen in Section 4.4.1.3.

However, we had reported effects specific to acoustic-models. Specially, FastPitch systems had displayed greater tilt lowering tendencies compared to Tacotron. This is not consistent in consonants, because differences betweeen the two acoustic models are not significant. Contrarily, we found consistent effects of the *vocoder*. First, Kruskal-Wallis test identifies significant group differences [χ^2(1)= 13.73, p-val $<$ 0.001]. Then, pairwise comparison shows that WaveGAN vocoders lower the tilt by 5.48 dB/log(Hz), and WaveNet by 4.31 dB/log(Hz). Systematic, individual system analysis reveals that vocoder-specific tendencies are consistent between each pair of systems that share an acoustic-model. Meaning, between Q and Y, Q lowers the spectral tilt more than Y. The same trend is available in R and Z.

Next, significant differences were found from the human voice on the basis of **spectral shape** [$\chi^2(1)$= 122.73, p-val < 0.0001]. This means that there is a greater difference between the spectral tilts above and below the mid-frequency range. Neural voices increase the spectral shape by 1.8 db/log(Hz) and 1.9 dB/log(Hz) in the CV and VC position, respectively. The effect is the clearest in the fricatives of the CV position. Here, spectral shape differences rise by 2.53 db/log(Hz) for voiceless fricatives and by 2.25 db/log(Hz) for voiced ones. No vocoder or acoustic model specific effects however, were observed.

4.5 Discussion

4.5.1 Impact of distortion on quality perception

In this chapter, we analyzed production characteristics of obstruents and their neighbouring vowels across different TTS systems. We found that neural voices deviate from the human voice the most in the context of voiceless fricatives. This observation suggests that consonants with a periodic source excitation are modeled more closely to the human voice than those with an aperiodic excitation. Our previous work (Pandey et al., 2022) was limited to WaveNet voices, and led us to hypothesise that the deviation emerged as a consequence of the autoregressive nature of vocoder. However, this reasoning is put into perspective especially when similar trends of deviation are seen in non-autoregressive synthesizers.

In the development of statistical parametric synthesis, there has been active interest in appropriately modelling excitation for unvoiced regions for high quality speech synthesis (Jensen et al., 1994; Drugman and Raitio, 2014). Parameters of voiced regions were estimated using periodic or quasi-periodic impulse trains, and unvoiced regions using white noise. To overcome the consequent buzziness, mixed excitation i.e., proportionally mixing the noise and periodic parameters gained popularity (Yoshimura et al., 2001, 2005; Yu et al., 2007).

In present-day modern synthesizers, supporting evidence comes from designs of Neural Source Filter Models (NSF) (Wang et al., 2019). Here periodic and aperiodic regions of the waveform are consciously modelled by separate source-filter combinations. Specifically, the h-NSF filter selectively allows high-pass filtering to the white noise excitation signal and merges the voiced/unvoiced components at the resultant waveform, instead of merging at the source signal. This improvement on the NSF results in comparable performance with WaveNet, and outperforms the baseline NSF. In other related work (Fujimoto et al., 2018) we find when WaveNet is enhanced with separate periodic/aperiodic decomposition it receives more favourable scores in naturalness.

However, an in-depth exploration of modelling voicelessness in end-to-end neural synthesizers is absent from the discussion on quality perception. As discussed in Section 4.2, obstruents suffer greater masking in impaired listening conditions and transmission (Li and Loizou, 2008, 2010). It is possible that feature distortion observed in neural TTS may cause information loss, that is not perceivable in ideal listening environments. Another consequence of improper segmental modelling can appear on low-resource settings where pre-trained models are frequently adapted (Prajwal and Jawahar, 2021; Debnath et al., 2020) to generate speech in a target, low-resource language. The adaptation is more challenging when the languages are not closely related. Pertinent to our case, the problem of *voicelessness* was identified in developing a Vietnamese code-mixed TTS, adapting a pre-trained English model (Nguyen et al., 2021). The mismatch occurred because Vietnamese does not support word-final fricatives as English does. While the use of speaker embeddings returned favourable naturalness, the problem remains open for other low-resource target languages.

Next, we found that the lowering of spectral tilt is a consistent trend in neural voices across all contexts, both in consonants and vowels. Previous studies have highlighted the importance of flatter spectral tilt on intelligibility (Lu and Cooke, 2009). Enhancing strongly negative tilts for voiced frames has resulted in improved naturalness and speaker similarity for synthetic speech (Sharma and Prasanna, 2017). Recent studies on masked speech also suggest a lowering of spectral tilt (Magee et al., 2020) results in a muffled speech output. Additionally, attributes such as pleasantness have been associated with energy in the high-frequency regions (VaroSanec-SkariC, 1999). Therefore, dampening high frequencies may result in degraded perception of voices.

4.5.2 Impact of distortion on contrastive perception

We have shown that contrastive features can be used to diagnose weaknesses of neural TTS synthesizers. This section provides an additional use of contrastive features, with applications of neural TTS in phonetics and speech science. Malisz et al (Malisz et al., 2019) argue that the high realism, naturalness by neural voices ensures that the neural TTS can now be used for phonetics research. In an additional, side-project (Pandey et al., 2023), we explored how phonemic contrast is maintained in neural TTS synthesizers. To investigate whether phonemic contrast is maintained in neural TTS in the same way as in a human voice, this paper provides **place classification** of English fricatives as a targeted test case. If contrast is encoded in the same parameters as in the natural voice, then TTS voices may be suitable tools for speech science research. Conversely, if phonemic contrast is indexed by divergent trends in TTS voices, then generalization may become more difficult. Additionally, unexpected acoustic detail may enforce a cognitive load condition and increase reliance on lexical cues (Mattys and Wiget, 2011). This can cause greater problems for non-native listeners (Mattys et al.,

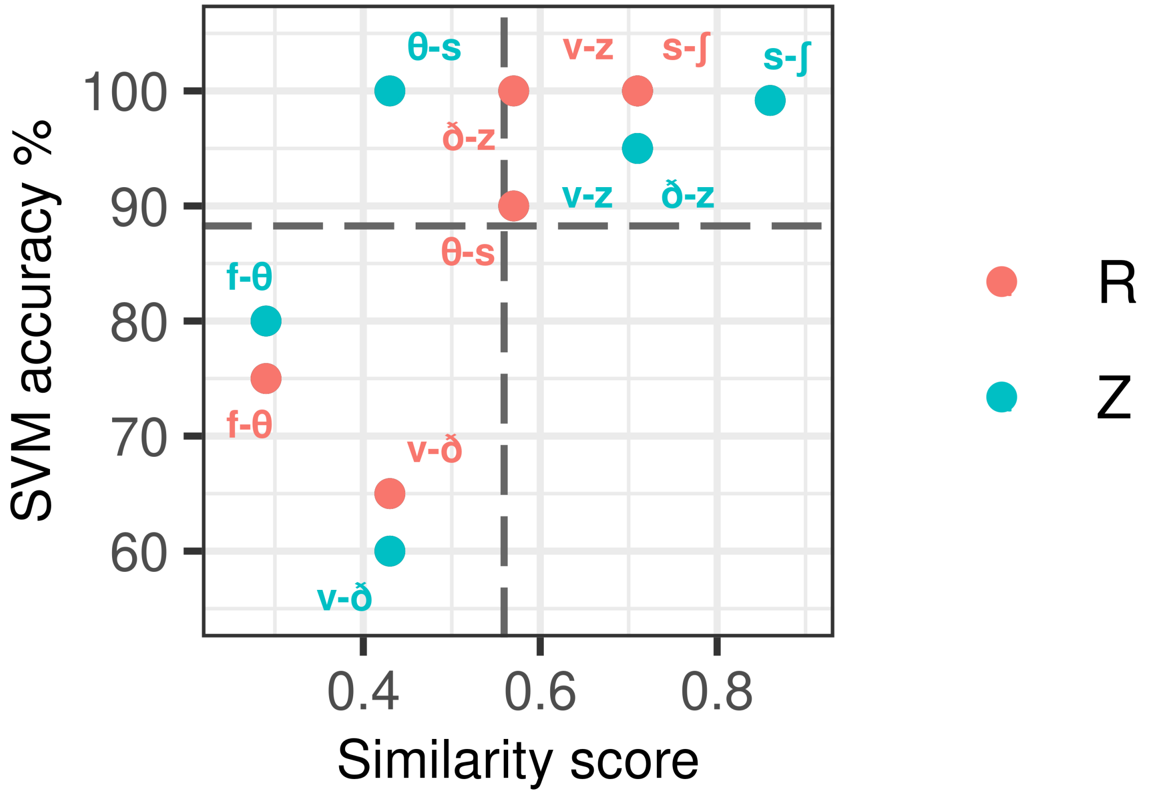


Figure 4.6: Relationship between accuracy and feature similarity with the human voice. Labels for contrastive pair presented for Tacotron WaveGAN (R) and Tacotron WaveNet (Z). Dashed lines show means on every axis.

2010).

For this, we conducted an analysis of feature importance using the voices of Tacotron WaveGAN and Tacotron WaveNet of the extended BC-2013. Tacotron voices were selected because more featural distortion was observed in the Tacotron voices (especially system R). Specifically, we compared the trends of place classification in human and Tacotron fricatives. 6 pairs of fricatives were classified using Support Vector Machines on the basis of contrastive features described in Section 4.3.2. Then, 7 most important features for the SVM classification was obtained using 1-AUC criterion through DALEX library in R (Biecek, 2018). Then, a Sorensen-Dice coefficient was used to compute a similarity score between the most important features in the human voice and each of the Tacotron voices. A high score meant that the contrastive trends were comparable between the human voice.

We found that sibilant fricatives produced by both Tacotron voices followed similar contrastive trends to the human voice. Figure 4.6 shows the concentration of sibilant fricatives in the top-right quadrant. This means that sibilant fricatives were classified accurately, and also used the same contrastive features for classification. Non-sibilant fricatives, on the other hand are shown in the bottom-left quadrant in Figure 4.6. This means that they differ from the human voice in the selection of important features, and that contrastive trends are not comparable. On the basis of these findings, we expect that the generalizations made on synthetic sibilants to extend to human sibilants. However, non-sibilants may require further investigation.

4.6 Conclusion

The MOS scores take us far enough to show that neural TTS is rated lower than the human voice. However, there is no explanation or diagnosis into the cause of this unnaturalness. In Section 2.4.2.4 we saw evidence from the studies where distortion or unnaturalness can be perceived at a microscopic level (Nusbaum et al., 1984; van Heuven and van Bezooijen, 1995), even when it does not result in conscious subjective ratings (Antons et al., 2012). We also saw in Section 2.4.2 and Figure 2.3 that present day evaluations do not consider a diagnostic analysis of the signal itself.

In a time-honoured quote, Ilse Lehiste writes that "*In the terminology of semantics, phonemes are signals not symbols.* (Lehiste and Peterson, 1959a)". This quote is relevant to the present discussion, because it ties together the idea of segmental evaluation using contrastive features of phonemes. The signal distortion in phonemic segments is used here to evaluate the speech synthesizer. Other metrics like PESQ also use a feature comparison between the human voice and degraded signal. However, contrastive features present characteristic information about a segment, which is expected to be robust to variations. Moreover, as described in Section 3.3.2, human listeners are more attuned to their perception.

This approach has several advantages over traditional evaluation of naturalness. First, it provides specific insights into the location of distortion. Since the deviation in voiceless segments and spectral tilt can be seen in 4 diverse neural architectures, a global tendency of neural TTS is revealed. This is particularly important for spoofing and fake-speech detection, especially in low-resource settings. Second, this has the potential of being developed as a full-reference and a no-reference metric for objective evaluation. A step-by-step outline is provided in Chapter 6. A final advantage caters to a more specific audience of speech scientists, and encourages their contribution to TTS research. This is because we draw directly from techniques in fundamental acoustic-phonetics and is automated to scale to large corpora. Since diagnostic trends are revealed in contrastive features, more segments (nasals, approximants) can be analyzed through this technique.

One final question remains: we do not know whether this distortion is perceivable to human listeners. This cannot be tested through complete utterances, as in MOS, because it is difficult to design complete utterances with obstruents. Moreover, we will not be able to disentangle the support of prosodic and contextual context with segmental distortion to clearly validate any trends. In the next chapter, design a subjective evaluation methodology, the *Long Arms* approach, which is more suitable to segmental distortion.

5 | Listener sensitivity to stimuli length and segmental distortion in WaveNet

5.1 Introduction

In the previous chapter, we saw that neural voices deviate from the human voice, especially for the voiceless regions. We also saw a blanket lowering of the spectral tilt in each of the neural voices. In this chapter, we explore the perceptual relevance of these effects on the perceived human-likeness of stimuli generated by the neural TTS synthesizers. Perception experiments are limited to WaveNet (i.e. Systems Y and Z), because both the WaveNet voices were ranked higher than the WaveGAN ones (See Chapter 3, Table 4.1). If segmental distortion is perceivable in higher-ranked systems, then WaveGAN systems can be extrapolated to sound distorted as well. The study involved presenting stimuli of varying lengths to 192 participants, who were asked to identify whether each stimulus was produced by a human or a machine. Their responses were captured using a 2-alternative forced choice task, and the results were analysed with a generalized linear model (GLM).

5.1.1 The logarithmic nature of human perception

Human perception is widely believed to be logarithmic (Fechner, 1860; Varshney and Sun, 2013; Dehaene, 2003; Ditz and Nieder, 2016). The non-linear responses to increases in stimuli was first observed by the nineteenth century philosopher, Ernst Weber. In measuring relative differences between perceived weights of objects, he identified that small changes in weight are more perceptible at the lower end of the scale (i.e, 20 g to 21 g). Conversely, in heavier stimuli (i.e, 40 g) the increase needed to be doubled, for the difference to be noticeable. This formulation was further advanced by his student, Gustav Theodor Fechner, who identified that the relationship between stimulus and perception was in fact, logarithmic. In other words, perceived change in sensation (ΔS) is proportional to the logarithm of the ratio between the stimulus intensity (S) and the reference stimulus intensity (S_0), with a constant of proportionality (k) that varies depending on the sensory domain. Together, this relationship was encoded as the Weber-Fechner law (Fechner, 1860), and is given by:

$$\Delta S = k \cdot \ln \left(\frac{S}{S_0} \right) \tag{1}$$

The Weber-Fechner law has been attested with evidence from a variety of sensory stimuli, such as taste, pressure, and intensity of brightness and amplitude (Varshney and Sun, 2013; Reichl et al., 2010). Consequently, we see a wide range of its commercial applications; such as in compression algorithms for maximizing visual comfort (Terzić and Hansard, 2016), identifying optimal sending rates in audio transmission (Chen et al., 2012), and even improving sweetness indices for artificial sweeteners (Mao et al., 2019). The Weber-Fechner law allows us to determine the extent of the change in stimulus, that will cause a perceptual response. To the best of our knowledge, it has not been introduced for synthetic speech evaluation. In this chapter, we evaluate if logarithmically varying stimuli can help in detecting machine-likeness in WaveNet stimuli.

Specifically, the influence of length is studied on the accuracy of responses. Participants are presented with stimuli of 2-syllables (e.g, "*gray wool*"), which increase in length by a multiplicative factor of 2, up to full-length utterances of 32-syllables (e.g, "*he ceased to satirize himself because time dulled the irony of the situation and the joke lost its humour with its sting*"). They are asked to rate these utterances simply as "human" or "machine", such that 2-alternative forced choice task (2AFC) captures their responses. With every doubling of stimulus length, we expect accuracy to increase with length. This is the *Long Arms* approach to subjective evaluation. Thus, the first question we explore in this Chapter is: *does the accuracy of human-machine detection increase with the length of the stimulus?* Next, in Chapter 4, we saw that features of voiceless obstruents, and spectral tilt deviates strongly from the human voice. If this deviation is perceivable, then utterances with segmental distortion should provide more clues as to the machine-likeness of an utterance. Hence, the second question we explore in this chapter is: *does accuracy of judgment increase even more when the stimulus is rich in obstruents?*

Three experimental conditions are designed to investigate these questions. Uniformly in each condition, stimuli vary logarithmically in length. In the first experimental condition, the **baseline** condition, the utterances are randomly selected from our dataset. In the second condition, the **obstruent-sonorant** or **ObSon** condition, obstruent-rich stimuli are compared with sonorant-rich ones. Finally, in the **spectral-tilt** condition, stimuli that deviate in spectral tilt are compared with those that do not. We found that utterances that are rich in obstruents are generally judged more machine-like, indicating that segmental distortion is perceivable. We also found robust perceptual differences between the two acoustic models in our study, i.e, Tacotron and FastPitch. The methodological details, results and discussion are described in the subsequent sections.

5.2 Experimental Design

5.2.1 Description of dataset

A complete description is provided of the dataset is provided in Section 3.2.1 and Section 4.3.1. A summarization is provided here for a standalone reading of this chapter. The source material for the experiments in this chapter comes from the recently extended (Le Maguer et al., 2022) version of the BC-2013 (King and Karaiskos, 2013). The human voice in the original challenge came from audiobook renditions by an American, female voice artist. The extended version (Le Maguer et al., 2022) contributes 4 neural voices, which are trained on the same human speaker as in the original challenge. Tacotron (Wang et al., 2017) and FastPitch (Łańcucki, 2021) were used as acoustic models for mel-spectrogram generation, and WaveNet (Van den Oord et al., 2016) and WaveGAN (Yamamoto et al., 2020) as vocoders for waveform-generation. However, only voices generated by the WaveNet vocoder are chosen: FastPitch WaveNet (System Y), and Tacotron WaveNet (System Z). The original human voice is constantly maintained as the reference.

All our stimuli were derived from the 100 utterances that originally formed the test corpus in both the original and extended versions of BC-2013. The next subsection explains the design and creation of the stimuli and presentation strategy.

5.2.2 Phrase extraction: text and audio

We developed a refined set of audio stimuli from the 100 utterances by taking the following aspects into consideration:-

Grammatical well-formedness:- First, we divided each utterance into its constituent phrases using the Stanford NLP parser (Manning et al., 2014). This ensured that our resultant phrases were grammatically well-formedness, and followed the syntactic structure of English. For example, we retained a syntactically appropriate noun phrase "*big, solemn oaks*", whereas a roughly cut up phrase *"before but she"* was discarded. This allowed our participants to focus only on the audio without getting sidetracked with grammatical anomalies. Finally, all duplicates were removed.

Phrase length:- After all the ill-formed phrases were pruned, phrases were further selected on the basis of length. The length of the phrase was determined in terms of the number of syllables per phrase. We only preserved unique phrases of *2, 4, 8, 16*, and *32* syllables to maintain a sufficiently perceivable “doubling" of their lengths. The number of phrases selected at each length is described in Table 5.1, as is the distribution of the stimuli between human and the two WaveNet systems. A total of 124 phrases was heard by each participant.

Phrase length (in #syllables)	#phrases	Human	FastPitch (Y)	Tacotron (Z)	Total phrases
2	64	32	16	16	**124**
4	32	16	8	8	
8	16	8	4	4	
16	8	4	2	2	
32	4	2	1	1	
Total	124	62	31	31	

Table 5.1: Number of phrases at each phrase-length heard by each participant across the human voice, and systems Y and Z.

Audio extraction:- The corresponding audio for the selected phrases was extracted from System Y, System Z and the human voice, and hand-corrected for phrases boundaries. Additionally, a fade of 50 ms was also added before and after each utterance, to minimise any audible clicks. The sampling rate was 44.1 kHz, bitrate 320 kbps, and the format was `.mp3`.

5.2.3 Experimental conditions and groups

As shown in Table 5.1, each participant evaluated 124 phrases. First, 62 human stimuli were extracted, in accordance with the phrase distribution in Table 5.1. These were maintained identically throughout the experiments. Synthetic stimuli were extracted based on one of the two conditions now described.

The baseline condition:- 62 synthetically produced stimuli of the required phrase length were selected randomly, with no particular constraints on their lexical content. This condition was designed to simply test the effect of increasing length on accuracy of human-machine detection.

The ObSon condition:- Each unique phrase among the well-formed phrases was assigned a score, based on the obstruent or sonorant concentration in its lexical content. Based on this score, phrases were categorized as obstruent-rich, or sonorant-rich[1]. Of the required 62 synthetic phrases, we selected 31 obstruent-rich phrases (OBS-P), and 31 sonorant-rich phrases (SON-P). This condition was designed for the second research question in Section 5.1.

[1] obstruent-rich: “mo**st** **s**e**lf** **p**o**ssessed**"; sonorant-rich:“**m**ea**ning** **in** it."

Based on Section 4.4, we expected the obstruent-rich stimuli to increase the accuracy of the human-or-machine responses.

The ObSon condition required us to also confirm whether the effect we hypothesized was truly an effect of obstruent-richness, and not that of a specific type of TTS system. In other words, we wanted to examine if this effect was consistent across both the acoustic models. Therefore, we designed our stimuli such that, for one group of participants, we retained those OBS-P which were produced by FastPitch (Y) and SON-P produced by Tacotron (Z). Then for another group, these pairings were reversed.

To maintain consistency between the two conditions, we also split the baseline stimuli equally, and paired them alternately with each of the acoustic models. But this split was completely arbitrary. The details of the individual participant groups are described in Table 5.2. Participants were assigned to one of the 4 groups listed. No participant was repeated in any group. Their details are described in Section 5.2.4.

The Spectral Tilt condition:- Segmental spectral tilt was calculated for the consonantal region, as well as at the transition boundary between the consonant and the vowel. Tilt estimates were calculated for every well-formed phrase, by averaging the segmental spectral tilt over the number of contributing segments, as described in Equation (2).

$$T_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \text{Segmental Spectral Tilt}_{i,j} \tag{2}$$

Here, T_i is the tilt estimate for the i^{th} phrase, and n_i is the number of contributing segments in the i^{th} phrase. Then, a phrase-by-phrase difference was calculated between the human voice and each system, as described in Equation (3).

$$\Delta T_i = T_i^{\text{human}} - T_i^{\text{system}} \tag{3}$$

Here, ΔT_i as the difference in tilt estimate for the i^{th} phrase between the natural voice and a particular system.

Phrases were ranked based on the magnitude of this difference. Those that deviated maximally formed the tilt-deviant (DEV-P) phrase set. Conversely, tilt-alike phrases (ALK-P) resembled the human voice in averaged segmental spectral tilt. Similar to the ObSon condition, we selected 31 phrases of each type, crossed them with the acoustic model, and assigned them in groups of participants.

Participant group	FastPitch (Y)	Tacotron (Z)	Human
Baseline_Group I	R1-P	R2-P	HM-P
Baseline_Group II	R2-P	R1-P	HM-P
ObSon_Group I	SON-P	OBS-P	HM-P
ObSon_Group II	OBS-P	SON-P	HM-P
SpTilt_Group I	ALK-P	DEV-P	HM-P
SpTilt_Group II	DEV-P	ALK-P	HM-P

Table 5.2: Phrases paired with acoustic model for each group. R1-P = Random Phrases 1; R2-P = Random Phrases 2; OBS-P = Obstruent-rich phrases; SON-P = Sonorant-rich phrases. ALK-P = Tilt-alike phrases; DEV-P = Tilt-deviant phrases.

5.2.4 Participant details

A total of 192 participants ($32_{participants}$ x $3_{condition}$ x 2_{group}) participants through Prolific (Palan and Schitter, 2018), an online participant recruiting platform. Gender balance was maintained for each group, such that each group contained an equal number of male and female participants. All participants were native speakers of English (UK or US English speakers, only), and reported no history of hearing impairment. Their informed consent was obtained prior to the experiment. The median time for completion was 25 minutes, and their remuneration rate was 7 GBP/hour. Quality control of crowd-sourced data was ensured through 4 attention checks. Here 4 audio samples of obvious machine-generated noise (e.g, sounds produced by a typewriter, coin collector, washing machine) were included in the evaluation stimuli. If participants marked the human-machine distinction correctly, their responses were recorded, otherwise discarded. However, we did not find any participants who failed our attention checks. Methods such as inter-annotator consistency (Graham et al., 2014) can also be used for ensuring the quality of crowd-sourced data.

The following demographic information was collected: a) age, b) sex at birth, c) speaker of UK or US English, d) experience with TTS devices or other TTS devices, and e) professional experience in speech/audio processing. We now display the within-group distribution of demographic information, to account for potential biases in their responses.

5.2.4.1 Distribution over participants' age

Figure 5.1 demonstrates the number of responses obtained from members of three broad age groups: young (18-35), middle-aged (35-50) and older (50-65) adults. As this was not a controlled study on participant ages, we can see that age-groups do not fall into neat thirds. Middle-aged adults are 1.8 times, while young adults are 2.78 times more numerous. This means that the largest proportion of our responses is contributed by young adults.

Sex is also unevenly distributed within some groups. Particularly, Group I in the Baseline

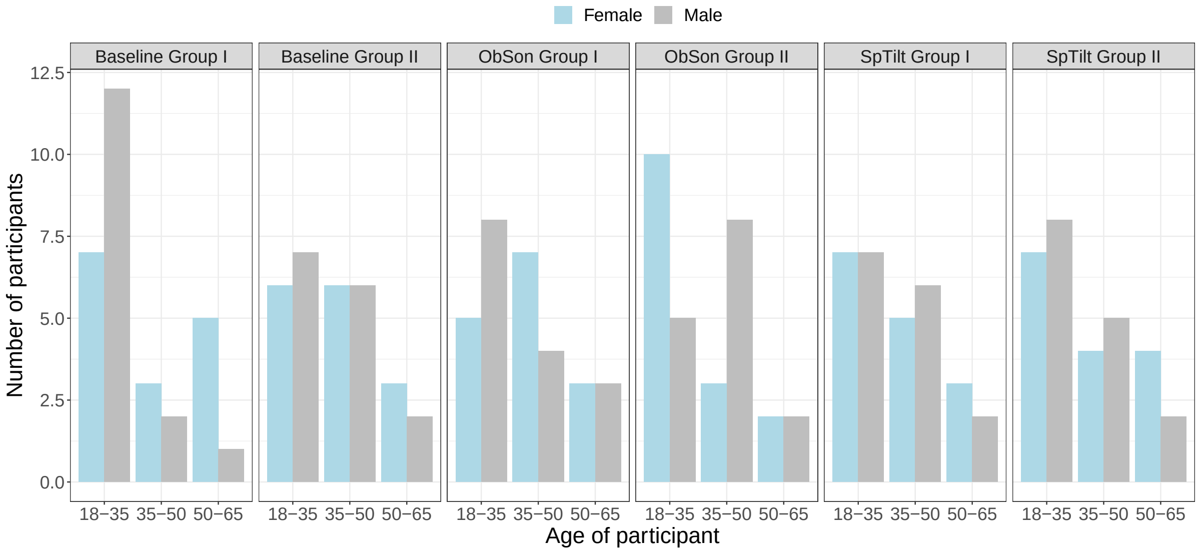


Figure 5.1: **Age-wise distribution:** Young, middle-aged and older participants in each group and experimental condition.

condition contains most responses from young male participants (Number of response: 1487_M, 867_F). Similarly, in the obstruent-sonorant condition, there are more young females than males, and vice versa in the middle-aged group. So it is important to note that, our data predominantly comes from younger and middle-aged participants. Therefore, we will only report broad trends in age-based differences.

5.2.4.2 Distribution over participants' exposure to TTS devices

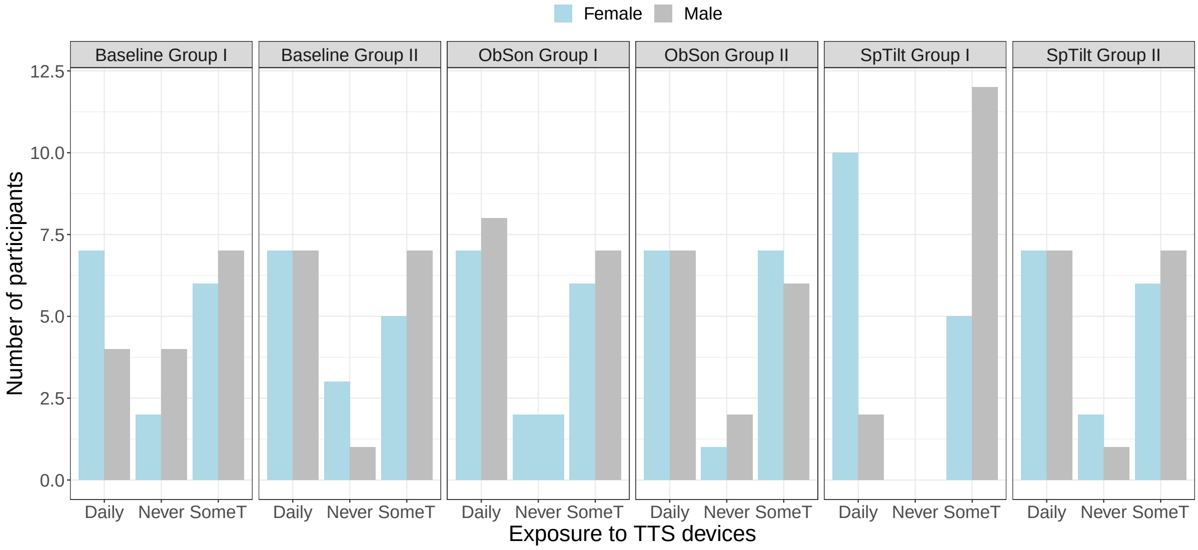


Figure 5.2: **Exposure-wise distribution:** Daily, Sometimes and Never users of TTS devices.

In Figure 5.2, we see the frequency distribution of participants based on their exposure to TTS devices. Participants self-report whether they are "Daily" users of Amazon Alexa, Siri etc, have an occasional exchange with them (coded as "Sometimes"), or have no exposure whatsoever to them ("Never"). Figure 5.2 shows a nearly uniform distribution between "Daily" and "Sometimes" users of TTS devices. This numerically motivates a robust statistical comparison between "Daily" and "Sometimes" users. However, the option "Sometimes" is somewhat vague, and the range of exposure cannot be fully determined. On the other hand, participants who report "Never" interacting with TTS provide a more concrete analysis variable for comparison with the "Daily" users. However, they are 4.4 times less numerous than "Daily" users. Particularly, Group II in the Spectral Tilt condition has no such participants, and Group I in the obstruent-sonorant condition has only 2. Finally, sex is approximately balanced between these subgroups. A bias is only observed Group I, Spectral Tilt condition, where there are more female "Daily" users, and male "Sometimes" users. Therefore, similar to the age demographic, broad trends in this data will be presented, with their generalizability subject to these confounds.

5.2.5 Presentation of the stimuli

The stimuli were presented in a random order, to remove any effect of sequencing on length. In every trial, we presented only one stimulus to the participant, and requested their response to the question: "*Did this sound like a human, or a machine?*". Stimuli were only played once. Their responses were captured in a 2-alternative forced choice task: "Human" or "Machine". The experiment was designed entirely in Psychopy (Peirce et al., 2022), and hosted online on the Pavlovia server [2]. The results of each experiment are discussed in Section 5.3.

5.2.6 Statistical model

Listener responses are coded as a binary variable where `0=wrong` i.e, the participant was wrong, and `1=correct` i.e, the participant is correct in detecting human or machine in a given stimulus. A logistic regression represents a sigmoid function, which accepts a real number t, as input and returns a value between 0 and 1 as its output.

$$\sigma(x) = \frac{1}{1 + e^{-t}} \tag{4}$$

x is the length of the stimulus, expressed as a continuous variable. For logistic regression, the input t to the sigmoid function is the linear combination of the following form:-

$$t = \beta_0 + \beta_1(x) \tag{5}$$

[2] https://pavlovia.org/

where, β_0 is the intercept and β_1 is the coefficient of regression. The sigmoid function, that predicts the likelihood of correct responses is given by:-

$$P(Accurate|stimulusLength) = \frac{1}{1 + e^{-(\beta_0 + \beta_1(stimulusLength))}} \tag{6}$$

The $P(Accurate|stimulusLength)$ is the likelihood that a response is correct, and will hereafter be referred to as P_{ACC}. A chi-squared test reveals the significance of the model, by comparing the full model to the intercept-only model. A p-value < 0.05 is considered statistically significant.

5.3 Results

Each subsection is dedicated to stimuli that share characteristics within an experimental condition. For example, obstruent-rich stimuli from the obstruent-sonorant condition are analyzed together. The sequence is as follows: Section 5.3.1 Human stimuli, Section 5.3.2 Randomly selected utterances, 5.3.3 Obstruent-rich Section 5.3.4 Sonorant-rich stimuli Section 5.3.5 Tilt-deviant and 5.3.6 Tilt-alike stimuli.

Next, system-specific effects are analyzed to identify those perceptual differences between the acoustic models of Tacotron and FastPitch, while the vocoder in each case was WaveNet. Reaction times and self-reported confidence scores are also presented, to identify whether increasing length ensures faster processing. Finally, we examine if different demographic variables, e.g. participants' age, exposure to TTS devices and sex, influence the chance of accuracy. Results of reaction times are expressed as a comparison of median values. All other results are analyzed through the GLM model, based on the likelihood of correct responses, i.e, P_{ACC}.

5.3.1 The human voice

In this section, we analyze utterances produced by the human speaker. Our aim is to observe whether our participants achieve higher accuracy as stimuli length increases. First, we discuss a combined model, where human stimuli from all the experiments are analyzed together. Next, we analyze data from each experimental condition (Baseline, Obstruent-Sonorant, Spectral Tilt), to identify any between-group effects among our participants. We supplement this analysis by presenting reaction times and confidence scores, and a discussion on demographic variables.

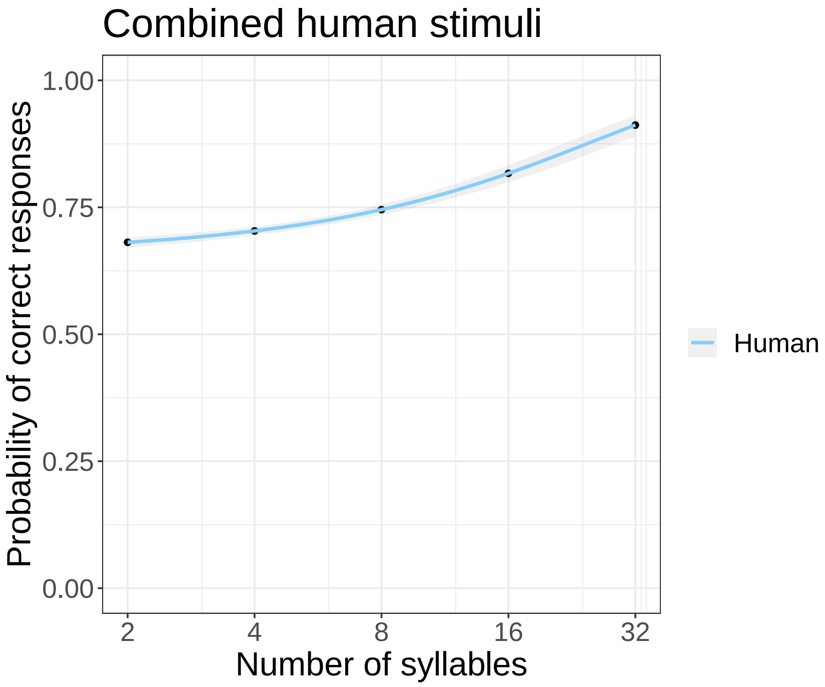


Figure 5.3: **Human stimuli:** A GLM-model fit combined over all the human stimuli across experimental conditions and participant groups. Model fit shows the relationship between the predicted probability of correct responses (0=incorrect, 1=correct) and increasing stimulus-length. Accuracy of correctly detecting human stimuli increases with stimulus length.

5.3.1.1 An overview of the human stimuli

First, participant responses on human stimuli are combined from all the experimental conditions to identify the general relationship between likelihood of correct responses, P_{ACC} and the stimulus length. Figure 5.4 shows a clear and gradual improvement in P_{ACC} with increasing stimulus length. This relationship is also strongly significant [slope (SE) 0.05 (0.004), p-val < 0.001]. At the shortest stimuli of 2-syllables, the P_{ACC} is 0.68, indicating that participants are likely to be correct 68% of times. At full-length utterances, this escalates 23.08% to 0.91. Comparing each individual increase in length, we find that P_{ACC} increases by 9.5% between the two longest stimuli, i.e. of 16 and 32 syllables. Similarly, a comparable spike of 7.14% are observed in the seond longest pair, i.e, between 8 and 16 syllables. A majority, i.e, 90% of participants provide correct responses for more than half of the 2-syllable trials. This high proportion is sustained, and further reaches 96.67% in full-length, 32-syllable utterances.

These results reflect that participants are highly likely to achieve correct responses on human stimuli, and that this likelihood increases with length. The next sub-section investigates whether this general trend is maintained over individual groups of participants.

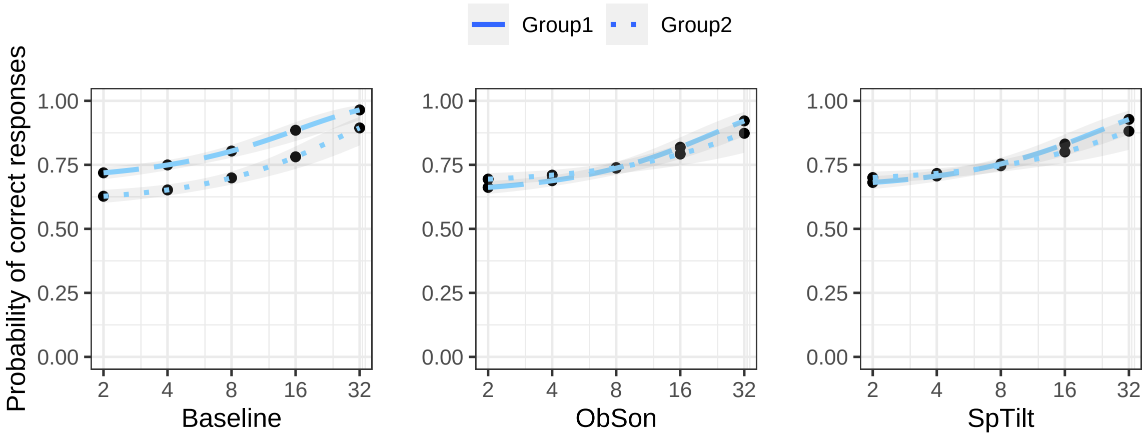


Figure 5.4: **Groupwise trends in human stimuli:** A GLM-model fit displaying data from individual experimental conditions. The fit shows the relationship between predicted probability of correct responses (0=incorrect, 1=correct) and increasing stimulus-length. Accuracy rises with length in all groups and experimental conditions.

5.3.1.2 Comparison between experimental conditions

Here, we individually examine the responses of participants in 3 experimental conditions, with 2 sub-groups within each. As discussed before, the lexical content of the human stimuli is identical throughout.

Length has a strongly significant influence [slope(SE) 0.06 (0.007), p-val < 0.001] on increasing likelihood of responses in the baseline experiment, where participants heard randomly selected WaveNet utterances. The relationship is consistent in both groups. Between the two endpoints, 2 and 32-syllable length stimuli, we see a 24.56% increase in Group I, and a 26.69% in Group II. Adjacent lengths are comparable across groups. Between 16 and 32 syllables, Group II however shows a sharper incline of 11.8%, compared to Group I at 9.5%. Finally, the proportion of participants who are at least 50% accurate, is somewhat higher in Group I. For instance, for long, 32-syllable utterances 100% participants are correct more than half the trials, while in Group II the proportion is 93.3%. In the obstruent-sonorant condition, increasing length holds a similar, strongly significant relationship with increasing P_{ACC} [slope (SE) 0.05 (0.007), p-val < 0.001]. The maximum P_{ACC} in Group I is 0.92, showing a 25.98% increase from shorter 2-syllable utterances. Although consistent in Group II, we see the maximum P_{ACC} numerically reduced to 0.87, and the endpoint difference to 17.98%. Comparing adjacent lengths, we see that the difference between each pair (2-4, 4-8..) is higher in Group I. This indicates a slightly faster rate of increase of P_{ACC} across lengths in Group I. Finally, the proportion of participants who provide 50% correct responses are comparable between the groups at most lengths, and identical at 96.67% for 32-syllable utterances. This means that, although the effects are consistent in both the groups, Group I has a faster rate of increase P_{ACC} as length of the stimulus increases.

Human stimuli in the spectral-tilt condition are also very similar to the combined model. P_{ACC} increases with increasing stimulus length, with strongly significant effects [slope (SE) 0.05 (0.007), p-val < 0.001]. P_{ACC} rises by 24.62% and 18.61% in Group I and Group II respectively. The maximum P_{ACC}, and the rate of increase is faster in Group I, as evidenced by a slightly sharper increase at every adjacent length. Like in other experimental conditions, a majority of participants (i.e, above 89.6%) in both groups score accurately on at least 50% of the trials, in every stimulus length.

The consistent high-accuracy across independently tested participant groups confirms no between-group effects on the perception of human utterances. Shorter stimuli are also rated above chance, showing that the distinction is clear with minimal input. This means, that participants can accurately perceive human-likeness in human speech, and that increasing length has a direct relationship with accuracy.

5.3.1.3 Reaction times and confidence scores

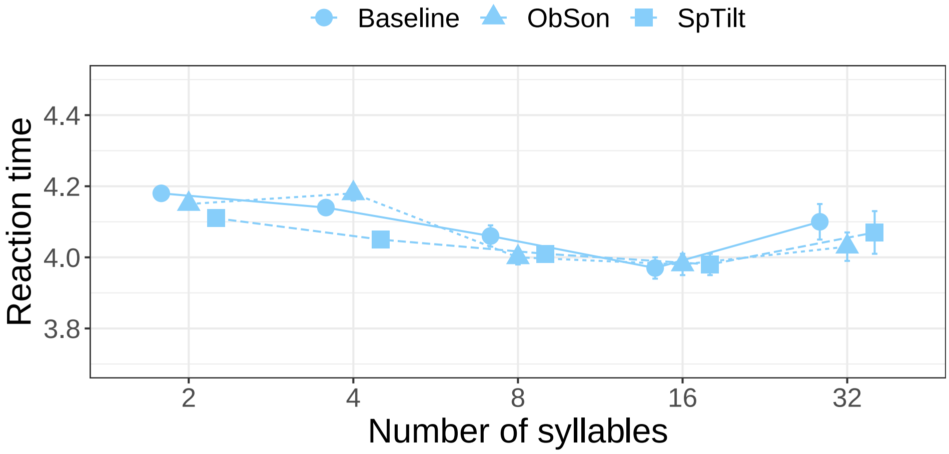


Figure 5.5: **Reaction times in human stimuli:** Data pooled from all participants. Error bars represent absolute deviation from the median, normalized by number of trials per stimulus length.

Reaction times fall consistently up to stimuli of length 16-syllables, and then increase at full-length utterances. As shown in Figure 5.5, this trend is uniform in all the experimental conditions.

Participants require a median of 4.14 seconds to rate 2-syllable utterances. In the baseline condition, participants are 80 ms slower than the spectral tilt, while the obstruent-sonorant condition is closest to the global median. The minimum reaction time is 3.97 seconds at 16-syllables, with negligible variability across experimental groups. At 32-syllables, this value increases by a median of 100 ms. Although consistent among conditions, baseline participants require the longest processing time of 4.10 seconds. Therefore, participants become gradually faster up to 16-syllable stimuli, but take more time to rate the full-length utterances. Since the lexical content of human stimuli was maintained identically, and we observe uniform between-group trends it is possible that a subset of utterances are responsible for this.

Next, the self-reported confidence scores increase with length across all experimental conditions. The proportion of participants who report "Very Confident" in their ratings rises by 35%, between short and full-length stimuli. The corresponding value is 75.7% in the Baseline condition, and 75.4% in each of the obstruent-sonorant and the spectral tilt condition.

Therefore, in rating human stimuli, participants become faster and more confident as length of the stimulus increases. This corroborates with the increased accuracy we had found in the previous section.

5.3.1.4 Influence of demographic variables

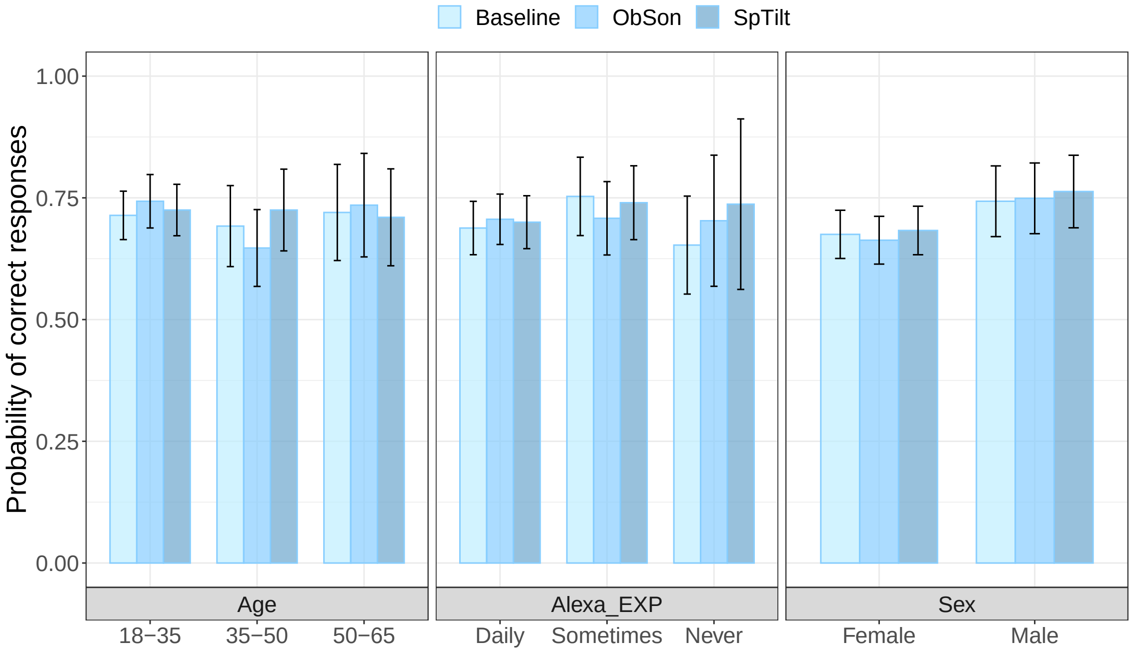


Figure 5.6: **Demographic variables in human stimuli:** influence of age, exposure to TTS and sex of the participant on the likelihood of correct responses P_{ACC} in each experimental condition. Error bars represent the deviation of the P_{ACC} about standard error of the logistic regression model.

Older adults do not differ significantly from younger adults. Middle aged participants show lower P_{ACC} (p-val < 0.001). But as seen in Figure 5.6, this is limited to only the obstruent-sonorant condition. In terms of exposure to TTS, occasional users have a 3.6% higher chance of providing correct responses. This effect is significant [slope (SE) 0.32 (0.08), p-val < 0.001] and consistent in the Baseline and the Spectral Tilt conditions. This means that participants who sporadically interact with TTS devices, have a higher chance of providing correct responses. On the other hand, participants who report no exposure to TTS devices do not differ significantly, or consistently among experimental conditions. Finally, as Figure 5.6 shows, male participants are, 8% more likely to respond correctly, with uniforma and statistically significant [slope (SE), p-val < 0.001] across experimental conditions.

Therefore, age does not have consistent effects on the chances of high accuracy. But higher likelihoods of accuracy are found in occasional users of TTS, and male participants in our population.

5.3.2 Randomly selected utterances

After a detailed discussion on human stimuli, we now analyze WaveNet stimuli in the Baseline condition. No prior consideration has been made for their selection, except the filtering described in Table 5.1. First, we combine data from both groups of participants. This helps us observe the general relationship between the length of stimulus, and the participants' chance of scoring correctly, i.e, the P_{ACC}. A system-specific comparison between FastPitch and Tacotron WaveNet is a recurring theme in subsequent sections. Through this, we explore system-specific perceptual patterns, which are often missed in MOS-based evaluations. Finally, demographic variables will be analyzed for their unique influence on participant performance.

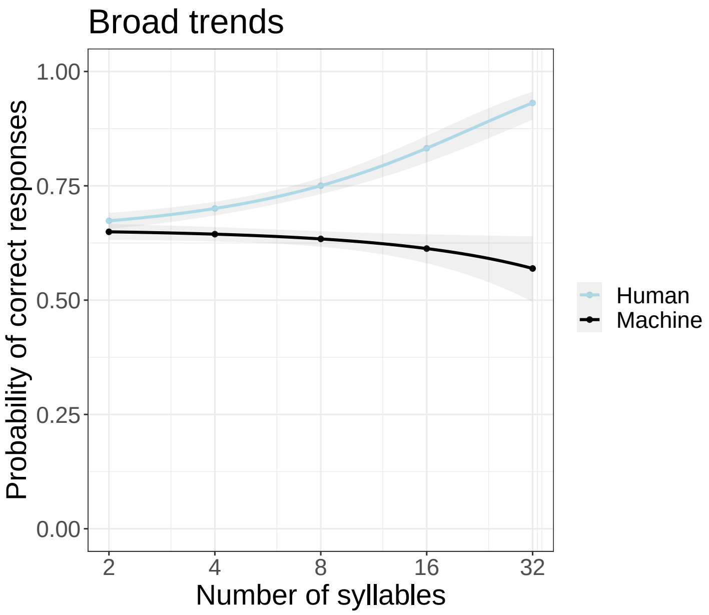


Figure 5.7: **Human vs Machine in the baseline condition**: A GLM-model fit comparing all human and machine stimuli from both groups and acoustic models in the baseline experimental condition. Model fit describes the relationship between the predicted probability of correct responses (0=incorrect, 1=correct) and increasing stimulus-length.

5.3.2.1 An overview of randomly selected stimuli

Participant responses are combined from both the groups, and stimuli from both the acoustic models in the Baseline condition. As Figure 5.7 shows, that P_{ACC} falls WaveNet, as opposed to the clearly rising ones seen in the human stimuli. A reduction of 8.15% in P_{ACC} is seen between the two endpoints, i.e. the shortest syllables of length 2 and full-length ones at 32 respectively. This trend indicates that participants find full-length utterances more human-like. Comparing adjacent lengths, we see that P_{ACC} falls with every incremental doubling. The sharpest drop of 4.12%, is observed between stimuli of length 16 and 32.

The proportion of participants who provide at least 50% correct responses shows an 8% increase at length 16, but does not increase at 32. Their average accuracy at length 2 is 62.8%, rises slightly to 69.6% and plummets to 45.83% at 32-syllables. These combined results suggest that human-machine distinction is more difficult in WaveNet utterances even with increasing length of the stimuli.

5.3.2.2 Comparison between Tacotron and FastPitch

Taken individually, Tacotron and FastPitch display trends that are consistent with the combined model, but vary uniquely in magnitude. The difference in individual behaviours can be seen more clearly in Figure 5.8.

In Tacotron, the length of the stimulus has a minor, but significant effect (slope (SE) -0.02 (0.01), p-val < 0.05) on lowering the likelihood of correct responses. Between the two end-points of length, Tacotron shows a 11.85% lowering of P_{ACC}. By contrast, FastPitch only lowers P_{ACC} by 4.16%, and the difference is non-significant. This indicates that participants make more mistakes with Tacotron utterances at longer lengths.

Groups were established to observe whether the effects were robust across lexical content. In Group I, FastPitch produces R1-P and Tacotron R2-P (see Table 5.1 for all details). A group-wise analysis (Figure 5.8 bottom panel) reveals divergent perceptual trends in each group. In addition to the acoustic model, this also points to the conditional effects of lexical content. In Group I, FastPitch shows a clear rising trend, such that full-length utterances are 13.27% more likely to be rated correctly, compared to the shortest ones. Conversely, in Tacotron, a steep drop of 26.84% is seen between the two endpoints. This drop is also strongly significant in Tacotron [slope (SE) -.04 (.01), p-val < 0.001]. Next, in the Group II, we know that phrase-sets are reverse-matched with the acoustic models. In FastPitch, which produced R2-P, the P_{ACC} shows a steep and significant fall [slope (SE) -0.03 (0.01), p-val < 0.01]. By contrast in Tacotron, length does not have a significant influence on increasing the participants' chance of being correct. P_{ACC} remains constant, with very minimal rises at every incremental doubling. It is only 3.08% higher at full-length utterances compared to the shortest, 2-syllable ones.

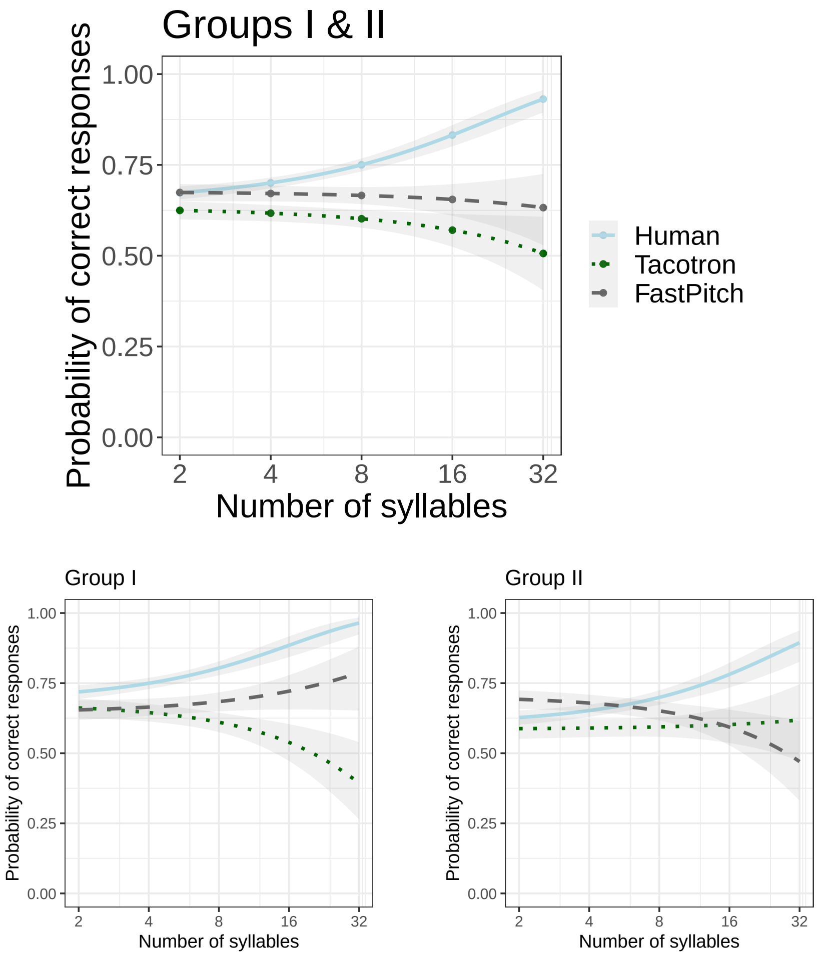


Figure 5.8: **Tacotron vs FastPitch in the baseline condition** A GLM-model fit showing differences in the baseline experimental condition. Model fit describes the relationship between the predicted probability of correct responses (0=incorrect, 1=correct) and increasing stimulus-length. The top panel shows data combined over both the groups, while the bottom panel shows individual groups. Trends in P$_{ACC}$ can be seen to vary with the lexical content of the phrase-sets for both the acoustic models.

Therefore, we make two important observations. First, when producing R2-P, FastPitch and Tacotron show similar trends of decreasing accuracy with longer stimulus lengths. However, when presented with R1-P, P_{ACC} either increases or remains unchanged. This clearly demonstrates that lexical content has an important influence on participants' ability to determine human/machine-likeness. Secondly, we observed that in FastPitch-produced R1-P, the rise in accuracy P_{ACC} is more pronounced (13.8%), compared to Tacotron (3.08%). In other words, participants have a higher chance of being accurate, and detecting machine-likeness in FastPitch for the same set of phrases. This suggests that there is a noticeable difference in the machine-likeness conveyed by each of these acoustic models.

More differences can be seen in the reception of the acoustic models, by analysing the proportion of participants who achieve at least 50% accuracy. This is lower in Tacotron at every length. At full-length utterances, while the proportion is 53.3% in FastPitch, it is only 38% in Tacotron. This means that a greater number of participants are more likely in detecting the machine-likeness of FastPitch produced utterances, compared to Tacotron ones.

Taking these results together, we find that Tacotron and FastPitch show measurable differences in their perceived machine-likeness. FastPitch utterances are rated more machine-like, while participants rate Tacotron as more human-like. Additionally, lexical content was also an important influence on the participant accuracy.

5.3.2.3 Reaction times and confidence scores

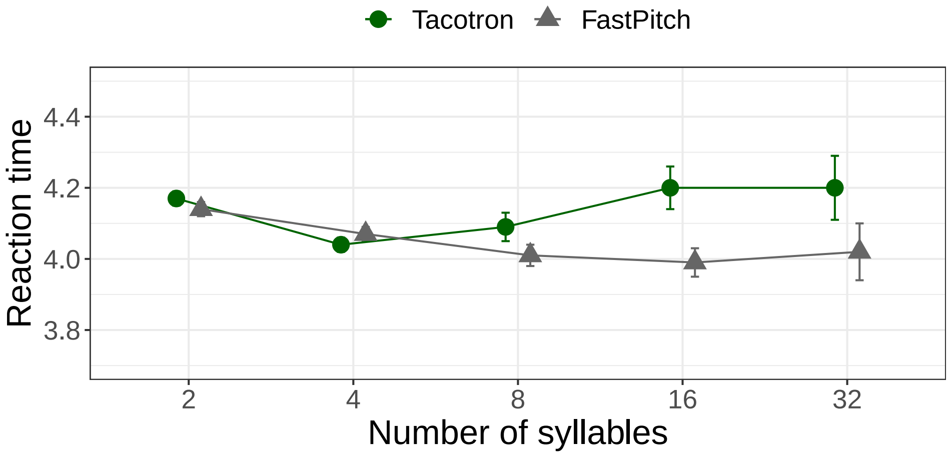


Figure 5.9: **Reaction times in the baseline condition:** Comparative display across acoustic models with data pooled from participants in both the groups. Error bars represent absolute deviation from the median, normalized by number of trials per stimulus length.

Reaction times vary between acoustic models, as seen in Figure 5.9.

In Tacotron, participants take an average of 4.17 seconds to rate the shortest utterances. This progressively decreases to 4.09 seconds for middle-sized, 8-syllable stimuli. However, participants take an additional 110 milliseconds, i.e., 4.20 seconds to rate utterances of 16 syllables and above. This pattern is consistent across both groups of participants, indicating that longer utterances in Tacotron require more processing time compared to shorter ones.

On the other hand, in FastPitch, reaction times generally decrease with increasing length. Participants take 4.14 seconds to evaluate 2-syllable stimuli, and the minimum reaction time of 3.99 seconds is recorded for 16-syllable stimuli. However, there is a slight increase of 30 ms in reaction times for full-length utterances. Upon further analysis, we identify that this is contributed by only one group (Group I), which also shows highest variance in the population. It is possible that a subset of participants are responsible for this increase.

The trends of self-reported confidence scores are similar in both acoustic models but differ in magnitude. In both models, stimuli of length 16 receive the highest confidence ratings. In Tacotron, 57.3% of participants report being "Very Confident" for length 16, while in FastPitch, this proportion is higher at 68.7%. For full-length utterances, the proportion is sustained in FastPitch, but drops to 46.6% in Tacotron. This indicates that fewer participants report being "Very Confident" in full-length utterances produced by Tacotron compared to FastPitch.

In summary, participants are generally quicker and more confident when rating FastPitch stimuli, which aligns with the higher accuracy and perceived machine-likeness discussed earlier. The next step is to analyze the demographic variables in our participant population to determine if the results are influenced by any particular group of participants.

5.3.2.4 Influence of demographic variables

An analysis of the demographic variables shows (see Figure 5.10) that they all uniquely influence the performance accuracy. The likelihood of providing correct responses declines with age, and is influenced by the sex and exposure of participants to TTS devices.

In terms of age-based differences older adults show significantly poorer performance than their younger and middle-aged counterparts. Older adults are 19.3% less likely to respond correctly, compared to younger ones. Age-based differences in both acoustic models are strongly significant (p-val < 0.001). Next, w.r.t exposure to TTS devices, participants with little to no experience with TTS devices show the best performance. They are 11.9% more likely to respond correctly, compared to daily users of TTS devices. Occasional users of TTS devices are also reportedly higher than daily users, especially in FastPitch (p-val < 0.001). Finally, sex-based differences vary in magnitude between acoustic models. Male participants are shown to have better performance, but the effect is stronger in Tacotron stimuli [slope (SE) +0.38 (0.09) p-val < 0.001), compared to FastPitch (p-val < 0.05).

Therefore, age-related differences can be seen, as older participants are less likely to be accurate in WaveNet stimuli. However, in terms of exposure to TTS, users who reported no

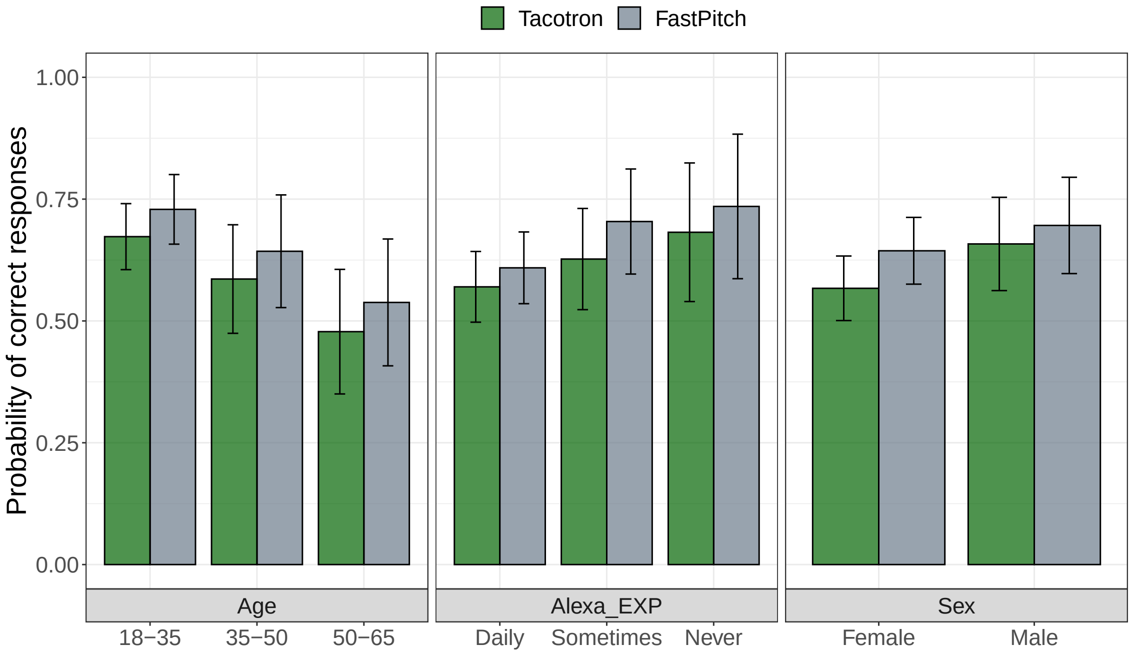


Figure 5.10: **Demographic variables in the baseline condition:** Influence of age, exposure to TTS devices and sex of the participant on the likelihood of correct responses P_{ACC} in each acoustic model in the baseline condition. Error bars represent the deviation of the P_{ACC} about standard error of the logistic regression model.

exposure to TTS devices were most accurate, followed closely by occasional ones. Daily users showed relatively poorer performance compared to these groups. Finally, male participants showed higher accuracy, especially for Tacotron stimuli.

5.3.3 Obstruent-rich stimuli

5.3.3.1 Overview of all obstruent-rich data

From Figure 5.11, it can be clearly seen that the probability of correct responses in obstruent-rich stimuli increases with increase in stimulus length. This indicates that the deviation in WaveNet obstruents, first reported in (Pandey et al., 2022), is perceptible and contributes to the perceived machine-likeness of the WaveNet stimuli.

Combined over both acoustic models, the P_{ACC} shows a sharp and consistent rise of 22.37% between stimulus length 2 and 32, when the stimuli are obstruent-rich. This difference is strongly significant [slope (SE) +0.03 (0.01), p-val < 0.001] P_{ACC} rises from 0.52 to 0.74, gradually ascending with every doubling in length. The sharpest effect of doubling can be seen between stimuli of lengths 16 and 32, where the jump is of 11.21%.

At longer lengths, we also find a larger number of participants who provide correct answers

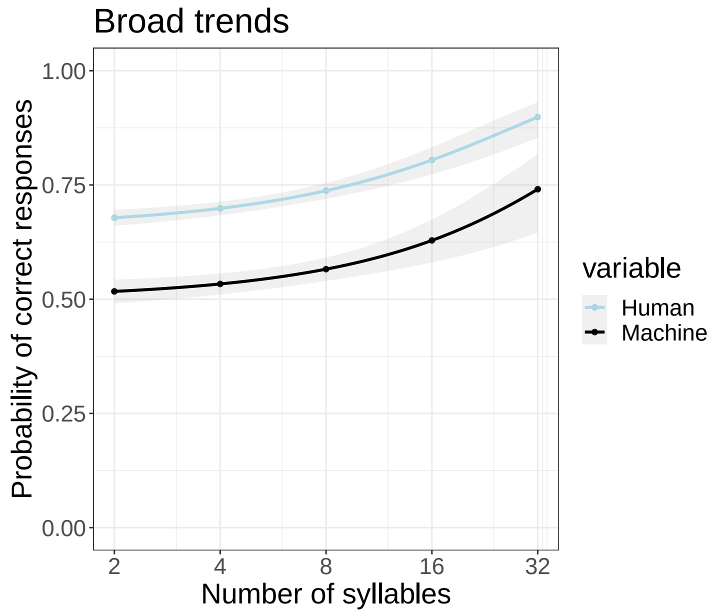


Figure 5.11: **Human vs Machine in obstruent-rich stimuli:** A GLM-model fit comparing human and obstruent-rich WaveNet stimuli, combined over both groups and acoustic models. Model fit describes the relationship between the predicted probability of correct responses (0=incorrect, 1=correct) and increasing stimulus-length. Generally, P_{ACC} can be seen rising for obstruent-rich machine stimuli.

for more than half the trials. While 63.2% of participants achieve at least 50% accuracy at shorter lengths (2-8), this value reaches 78.3% and 81.7% at lengths 16 and 32 respectively. Their average accuracy at length 2 is 51.2%, which rises to 81.7% for stimulus length 32.

Therefore, when the stimuli are obstruent-rich, then the likelihood of correct responses, the average accuracy and the proportion of participants achieving at least 50% correct responses, increase with stimulus length.

5.3.3.2 Comparison between Tacotron and FastPitch

There are notable distinctions between the voices generated by the Tacotron and FastPitch acoustic models. As in Figure 5.12, FastPitch exhibits a more pronounced increase in accuracy as a function of length, compared to Tacotron. Between stimuli lengths 2 and 32, the P_{ACC} of FastPitch increases by 26.65%, while Tacotron by 17.85%. This indicates that participants are more likely to classify FastPitch-generated stimuli as machine-like than those produced by Tacotron. Comparing adjacent lengths, we find P_{ACC} shows sharper rises between adjacent stimuli, while in Tacotron the movement is gentler. Between 16 and 32, the P_{ACC} in FastPitch rises by 12.71%, and in Tacotron by 9.24%. However, between 8 and 16, FastPitch shows another spike, as it rises by 7.75%. In contrast, Tacotron shows a softer rise of 4.89% between 8 and 16. This indicates that machine-likeness in FastPitch is more easily detectable at relatively shorter stimuli lengths.

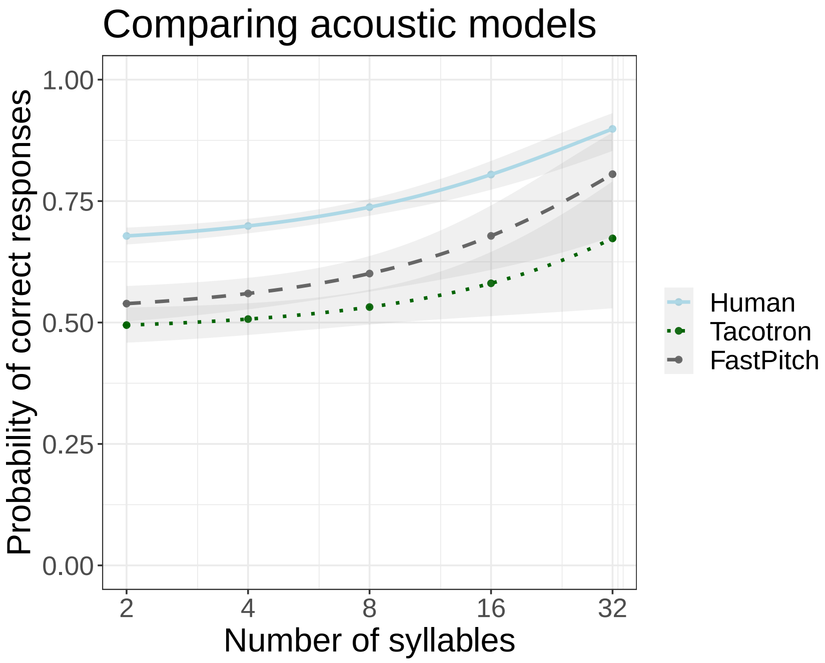


Figure 5.12: **Tacotron vs FastPitch in obstruent-rich stimuli:** A GLM-model fit showing differences between the two acoustic models in the obstruent-rich stimuli. P_{ACC} rises for obstruent-rich stimuli for both groups and acoustic models. But a faster and sharper rise can be seen in FastPitch.

Similarly, trends of participants who achieve at least 50% accuracy are comparable at shorter lengths, but differ more when the stimulus length increases. At length 2, this proportion is comparable between systems, i.e, 63.3 % in FastPitch, and 60% for Tacotron. But at length 32, FastPitch escalates to 86.7%, while Tacotron shows a more modest rise up to 76.6%.

Therefore, while obstruent-rich stimuli are overall judged to be machine-like, the accuracy is higher, and rises faster for utterances produced by FastPitch.

5.3.3.3 Reaction times and confidence scores

As can be seen in Figure 5.13, reaction times fall considerably with increasing phrase-length (see Figure 5.13. This trend is uniform in obstruent-rich utterances produced by both the acoustic models, but notably sharper in Tacotron.

In Tacotron, participants require a median of 4.10 seconds to rate 2-syllable utterances. Progressively declining, this value lowers by 220 ms, and hits a minimum of 3.88 seconds for full-length utterances. This demonstrates that participants are faster at responding to the Tacotron stimuli.

Similarly in FastPitch, we see a gradual reduction in reaction time, indicating faster processing for longer obstruent-rich utterances. Comparable to Tacotron, shorter, 2-syllable utterances require 4.11 seconds. At full-length utterances, we record a median of 4.01 seconds. Although numerically higher than Tacotron, this group also reflects greater variance in the participant population.

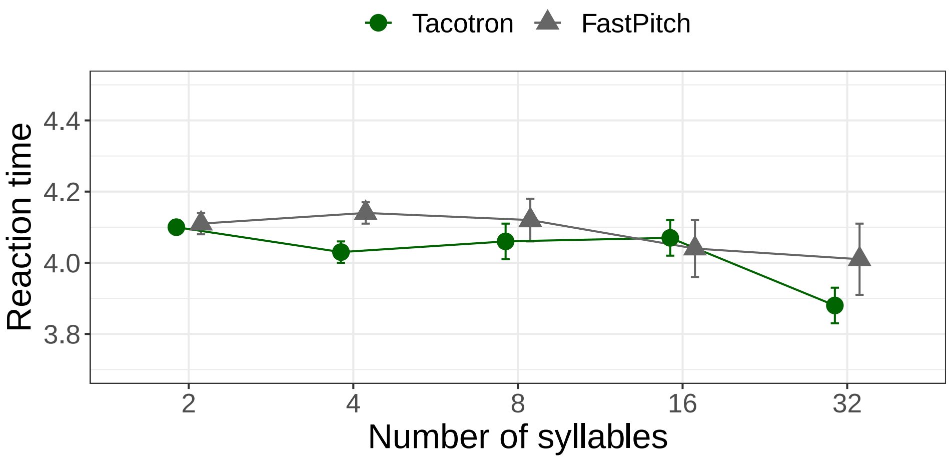


Figure 5.13: **Reaction times in obstruent-rich stimuli:** Comparative display across acoustic models with data pooled from participants in both the groups. Error bars represent absolute deviation from the median, normalized by number of trials per stimulus length.

Finally, self-reported confidence scores show further differences. For shorter, 2-syllable utterances, only 3̃5% of participants report "Very Confident". However, as length increases, this proportion escalates to 86.2% in FastPitch, but remains at 58.6% in Tacotron.

Taking all these results together, we infer that although participants are quicker to judge Tacotron produced stimuli to be more machine-like, fewer participants report high confidence about their judgement.

5.3.3.4 Influence of demographic variables

Analysing the influence of demographic variables, we find that age, sex and experience with TTS devices uniquely influence the probability of achieving correct scores.

Compared to younger adults (aged 18-35), older adults are 19.5% less likely to respond correctly. This difference is consistent and significant across both acoustic models, with sharper effects in Tacotron [slope (SE) -0.65 (0.20), p-val < 0.001]. This means that although older adults are overall likely to respond incorrectly, they are even poorer with Tacotron. Middle-aged adults also responded less accurately, but the difference is not significant. In terms of exposure to TTS devices, participants who report having no experience with TTS devices show the worst performance, with modestly significant differences. Compared to daily users of TTS devices, they were 16.6% less likely to respond correctly. This age-based difference is strongly significant in FastPitch [slope (SE) -0.90 (0.04), p-val < 0.001], but minimally so in Tacotron. As can be seen in Figure 5.14, daily users also score lower for Tacotron, narrowing the difference

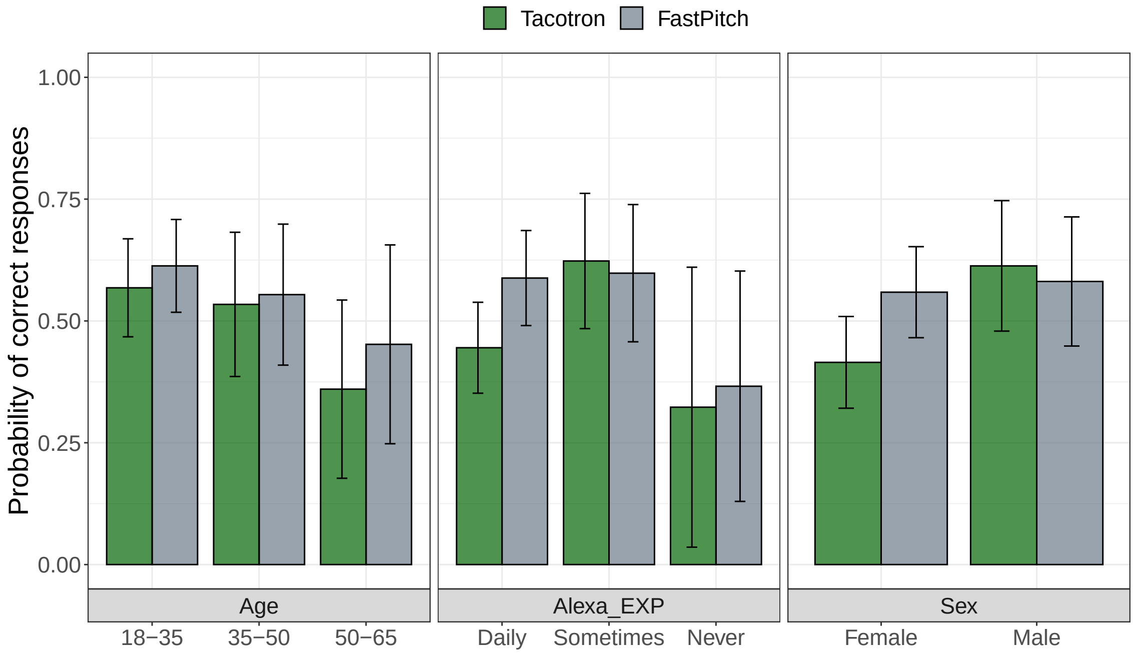


Figure 5.14: **Demographic variables in obstruent-rich stimuli:** Influence of age, exposure to TTS devices and sex of the participant on the likelihood of correct responses P_{ACC} in each acoustic model for the **obstruent-rich** WaveNet stimuli. Error bars represent the deviation of the P_{ACC} about standard error of the logistic regression model.

between them and no-exposure users. This indicates that regular exposure can aid human-machine detection only in some cases. Occasional users, on the other hand, show maximum chances of being correct. Finally, in terms of sex based differences, male listeners are 11% more likely to respond accurately. This difference is not significant in FastPitch, but strongly significant in Tacotron. As Figure 5.14 shows, females are less likely to be accurate for Tacotron stimuli.

Therefore, analysing influence of other variables on the accuracy of participants we find that older adults, and those who report no exposure to TTS devices show considerably lower accuracy to younger ones, and daily and occasional users of TTS devices respectively. Consistent with other experimental conditions, male participants are more likely than female ones to respond correctly, especially in Tacotron. Effects are mostly uniform across both the acoustic models, but females and daily listeners of Tacotron show comparatively lower P_{ACC}. This points to an influence of Tacotron stimuli over specific sub-groups.

5.3.4 Sonorant-rich stimuli

After discussing randomly selected and obstruent-rich stimuli, this section analyzes those phrases which are rich in sonorants. Figure 5.15 shows the combined model, plotted over both the acoustic models across both groups of participants. We can see a flat slope, conveying

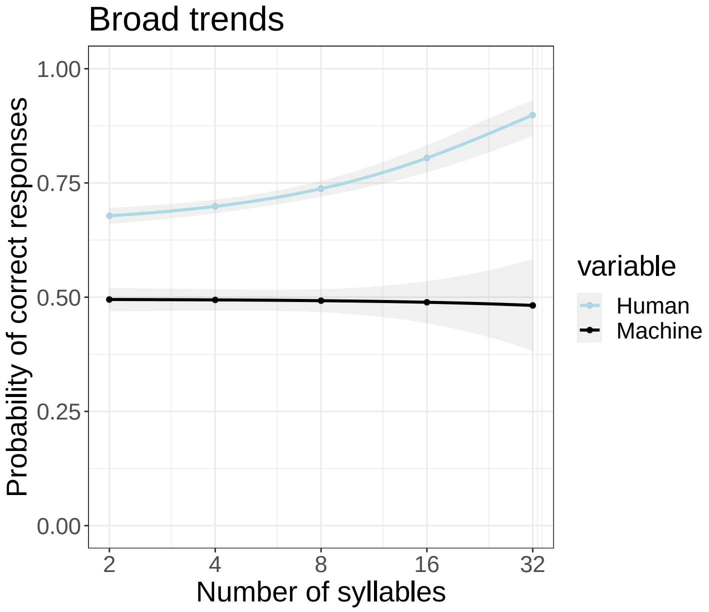


Figure 5.15: **Human vs Machine in sonorant-rich stimuli:** A GLM-model fit comparing human and sonorant-rich WaveNet stimuli, combined over both groups and acoustic models. Model fit describes the relationship between the predicted probability of correct responses (0=incorrect, 1=correct) and increasing stimulus-length. Generally, P_{ACC} can be seen unaffected with stimulus length for sonorant-rich stimuli.

that length does not have a significant influence on the chance of participants' scoring higher. It is clear that the trends are quite different from the obstruent-rich data.

5.3.4.1 Overview of all sonorant-rich data

Although not statistically significant, we find that accuracy is likely to fall as length increases. There is a small lowering of 1.98% in P_{ACC}, between stimuli lengths 2 and 32, i.e, the endpoints of stimulus length. This means that participants are slightly less likely to respond correctly for full-length utterances, than they are to shorter ones. Next, comparing other adjacent lengths, we find that P_{ACC} lowers by <1% at every doubling of stimuli length. For example, between stimuli of 16 and 32 syllables, where we expect maximum differences, we see only 0.69% lowering.

Next, participants who achieve accuracy for at least half of the trials is also relatively constant across lengths. This value is 56.7% at length 2, and only rises up to 61.7%. This increase is much smaller, compared to a 20% rise in obstruent-rich stimuli.

Therefore, length does not appear to aid perception of machine-likeness in sonorant-rich stimuli. However, now we explore if these trends are motivated by a particular acoustic model, or participant group, and whether they are reflected in reaction times.

5.3.4.2 Comparison between Tacotron and FastPitch

The system-specific analysis shown in Figure 5.16 reveals clear trends of perceptual differences between the two acoustic models, FastPitch and Tacotron. Full-length, sonorant-rich utterances produced by FastPitch are more likely to be detectable as machine-like, while the falling accuracy may be contributed by Tacotron.

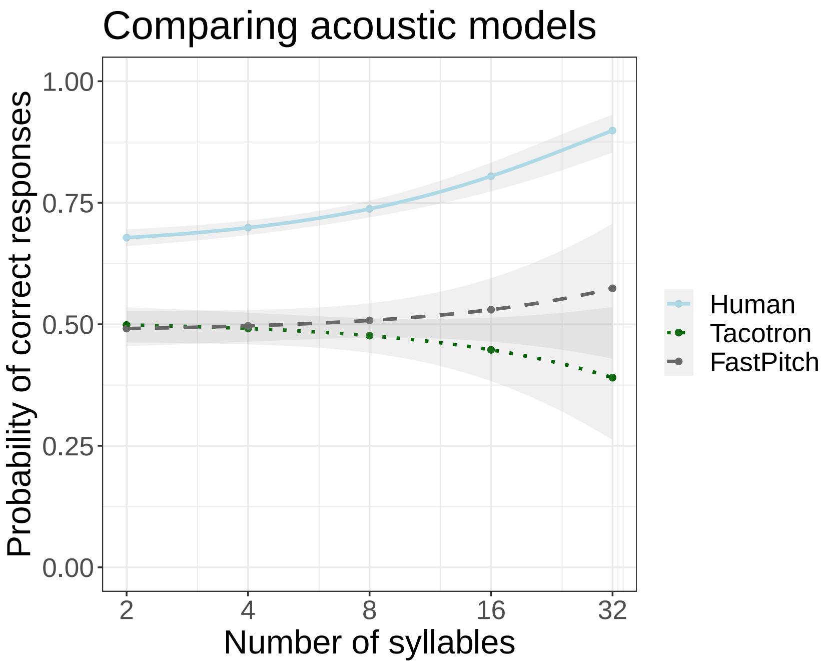


Figure 5.16: **Tacotron vs FastPitch in sonorant-rich stimuli:** A GLM-model fit showing differences between the two acoustic models in the sonorant-rich stimuli. P_{ACC} rises for FastPitch but falls for Tacotron.

Tacotron stimuli show a consistently falling pattern, indicating that the stimuli sound more human-like with every increase in stimulus length. Between the two endpoints, P_{ACC} falls by 10.84%. Upon comparing adjacent increments, we find gradually falling P_{ACC} with every doubling. The sharpest decline of 5.72% is seen when length doubles from 16 to 32. In FastPitch, these trends are reversed. Sonorant-rich stimuli show a small but increasing trend in accuracy, showing that machine-likeness of these stimuli is perceivable. There is 8.27% increase in P_{ACC} between the shortest and full-length utterances, showing that an increase in length aids the perception of machine-likeness. In the same vein, 76.67% of participants in FastPitch obtain correct responses for at least half of the trials, while that proportion drops to 50% in Tacotron in full-length utterances.

Taking these results together, we find that sonorant-rich stimuli sound machine-like when produced by FastPitch, but appear human-like when by Tacotron. These trends are consistent with our hypothesis, but not statistically significant. So, the predictive strength of this analysis is somewhat limited for future results.

5.3.4.3 Reaction times and confidence scores

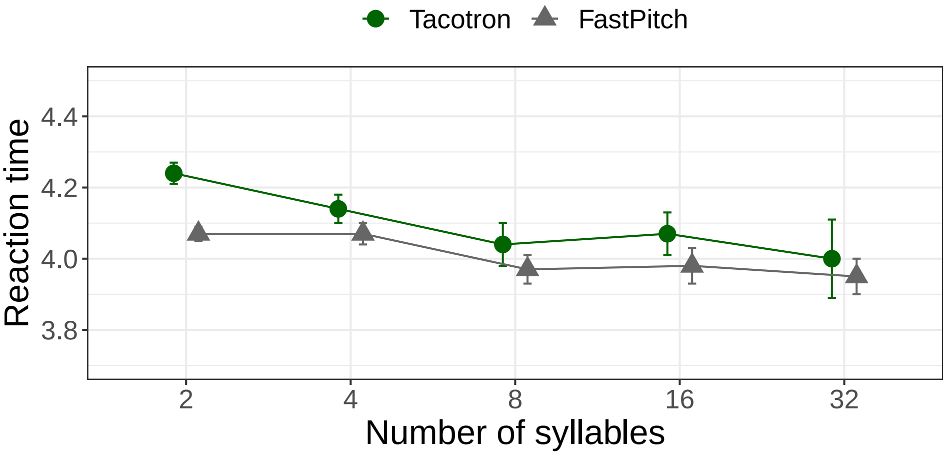


Figure 5.17: **Reaction times in sonorant-rich stimuli:-** Comparative display across acoustic models with data pooled from participants in both the groups. Error bars represent absolute deviation from the median, normalized by number of trials per stimulus length.

In sonorant-rich stimuli, especially for FastPitch, participants are overall faster at decision making with increasing stimulus length.

In Tacotron, participants require a median of 4.24 seconds to judge the shortest, i.e, 2-syllable stimuli. At 32-syllables, the reaction times drop by 240 ms and reach 4.00 seconds. However, at 16-syllables, we see an inconsistent increase, with expected variance in the participant population. This points to an inconsistent relationship of Tacotron with length in sonorant-rich stimuli. On the other hand, in FastPitch, we find that reaction times consistently fall. They are highest at 4.07 seconds, and lower quickly to 3.97 at middle-length, i.e 8-syllable utterances. This value is sustained as length increases to 32-syllables. So, comparatively, FastPitch enables a faster processing of shorter utterances, while Tacotron prompts the participant to continue incorporating information from exposure.

Finally, higher confidence is associated with FastPitch at every length. The proportion of participants who report "Very Confident" in their ratings rises from 35.4% to 69.0% in FastPitch. In Tacotron, this proportion moves between 26.08% to 57.1%.

Recall from Table 5.1, that non-overlapping groups of participants rated sonorant-rich stimuli from each acoustic model. Specifically, Group I rated FastPitch, and Group II rated Tacotron. Our results from within-group analysis support that sonority of the lexical content has a influence on confidence scores. For example, the proportion of participants who are "Very Confident" drops by 10.4%, and 29.1% in Group I and Group II respectively. The sharper difference in Group II, further highlights the contrast between FastPitch produced obstruent-rich and Tacotron produced sonorant-rich ones. In other words, participants detect machine-likeness confidently in FastPitch obstruents.

5.3.4.4 Influence of demographic variables

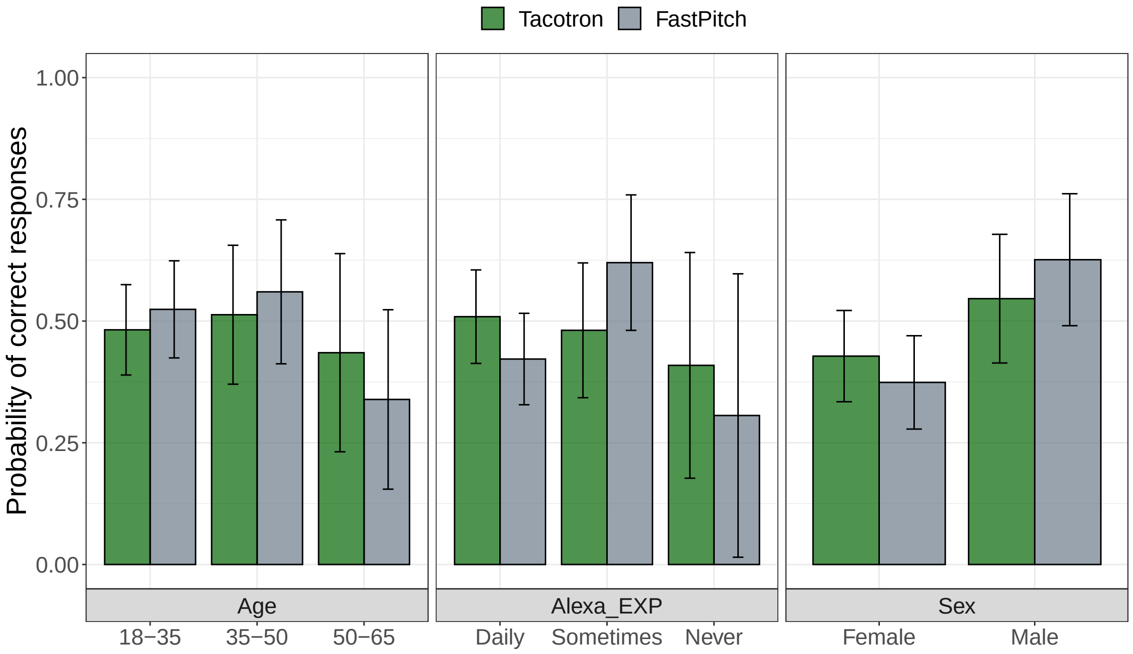


Figure 5.18: **Demographic variables in sonorant-rich stimuli:** Influence of age, exposure to TTS devices and sex of the participant on the likelihood of correct responses P_{ACC} in each acoustic model for the sonorant-rich WaveNet stimuli. Error bars represent the deviation of the P_{ACC} about standard error of the logistic regression model.

As in obstruent-rich stimuli, we find that age, sex and experience with TTS devices contribute individual influences to the probability of achieving correct scores.

Compared to younger adults (aged 18-35), older adults are 12.4% less likely to respond correctly. However, the difference is only significant in FastPitch [slope (SE) -0.76 (0.18), p-val < 0.001], and not in Tacotron. Middle-aged adults do not show significant differences in either acoustic model. Next, compared to daily users of TTS devices, participants who report having no experience are 9.6% less likely to respond correctly. This difference is only minimally significant across both Tacotron and FastPitch (p-val < 0.1). On the other hand, responses of occasional users are not consistent across acoustic model. In FastPitch they display a strong, and statistically significant likelihood of being accurate [slope (SE) 0.81 (0.14), p-val < 0.001], but not specially different in Tacotron. Lastly, sex-based differences are consistent with previous trends. Male listeners are 18.6% more likely to respond accurately. This difference is also statistically significant (p-val < 0.001) in both acoustic models.

It must be noted that sonorant-rich FastPitch utterances were rated by Group I participants. As noted before, those who reported no exposure to TTS devices were also female participants in the older age group. It is possible that the human-like reception of FastPitch is biased by these participants, and the obtained P_{ACC} is lower. However, we refrain from an in-depth analysis here.

5.3.5 Spectral tilt: tilt-deviant stimuli

We now move to the third experimental condition, where the stimuli, either deviant or alike, are presented to our participants. This section analyzes stimuli produced by WaveNet, that differ from the human voice in segmental spectral tilt.

5.3.5.1 Overview of all stimuli that deviate in spectral tilt

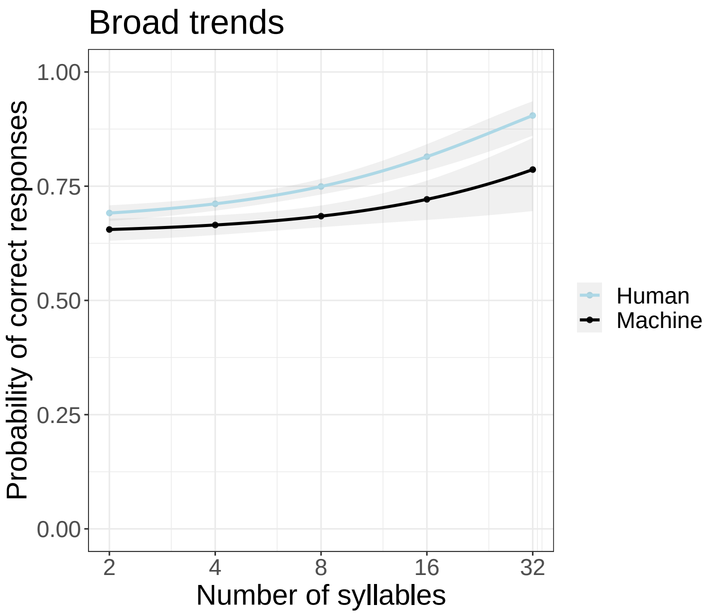


Figure 5.19: **Human vs Machine in tilt-deviant stimuli:** A GLM-model fit comparing human and tilt-deviant WaveNet stimuli, combined over both groups and acoustic models. Model fit describes the relationship between the predicted probability of correct responses (0=incorrect, 1=correct) and increasing stimulus-length. Generally, P_{ACC} can be seen rising for tilt-deviant machine stimuli.

From Figure 5.19, we can see that the chance of providing correct responses steadily increases with length. P_{ACC} rises from 0.66 to 0.79 between the endpoints of stimuli length, i.e, 2 and 32. This 13.12% rise is a statistically significant [slope (SE) 0.02 (0.01), p-val < 0.01]. It must be noted that even for shorter stimuli, the P_{ACC} is fairly high compared to other stimuli we have seen before. The sharpest rise is between stimuli of length 16 and 32, where we see a 6.51% increase in P_{ACC}.

Most participants score accurately, as is evidenced by the proportion where they provide at least 50% correct responses. At stimulus length 2, this proportion is high at 75.4%, while it reaches 85.25% for full-length utterances. Their average accuracy is 63.3% at shorter stimuli and peaks at 85.2% for full-length utterances. These trends together indicate that stimuli that deviate from the human voice in terms of spectral tilt are likely to be judged machine-like. The following sections explore the individual effects of and acoustic models, participant groups and demographic variables.

5.3.5.2 Comparison between Tacotron and FastPitch

Taking individual systems apart, we find variable trends in the perception of acoustic models. FastPitch stimuli rise sharply, comparable to an increase in the human stimuli. This indicates that machine-likeness is as clear in tilt-deviant FastPitch, as human-likeness is in human stimuli. Notably, shorter stimuli in both acoustic models retain their high P_{ACC}, indicating that the tilt-deviant stimuli are fairly clear in machine-likeness even with different participant groups.

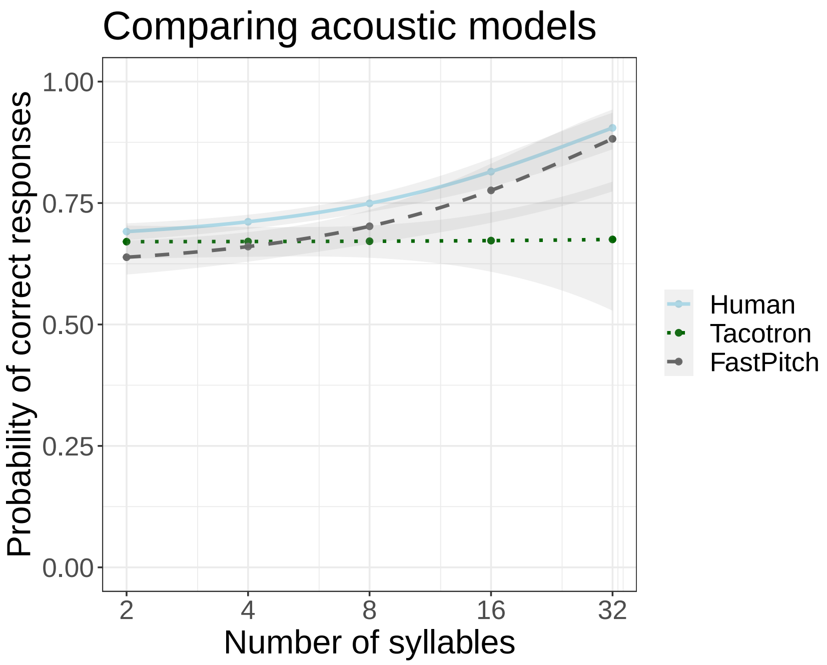


Figure 5.20: **Tacotron vs FastPitch in tilt-deviant stimuli:** A GLM-model fit showing differences between the two acoustic models in the tilt-deviant stimuli. P_{ACC} rises for FastPitch but remains more-or-less constant for Tacotron.

Stimuli length aids the chance of correct responses by 24.47%, and is strongly significant in FastPitch [slope (SE) 0.05 (0.01), p-val < 0.001]. Comparing adjacent stimuli, we see a maximal rise of 10.32% between stimuli of length 16 and 32. The next longest pair of stimuli, i.e, 8 and 16 also show a fast incline of 7.39%. Conversely in Tacotron, we do not see a statistically significant effect of length. The endpoint difference between stimuli lengths 2 and 32 is only 0.46%, indicating little support from length in machine-like perception. At other adjacent lengths, P_{ACC} remains unchanged, with relative differences not exceeding 0.1%.

Participant data retains its high accuracy from the combined model. Participants who provide correct answers at least 50% of the time are above 70% in both Tacotron and FastPitch. This proportion further increases to 90.32% in FastPitch, while remains sustained at nearly 80% in Tacotron. Most participants, therefore, score correctly on half or more trials.

Taking these results together, we can clearly see differences in the relative perception of acoustic models. Now let us explore the trends of reaction times, and the unique influence participant demographics on these trends.

5.3.5.3 Reaction times and confidence scores

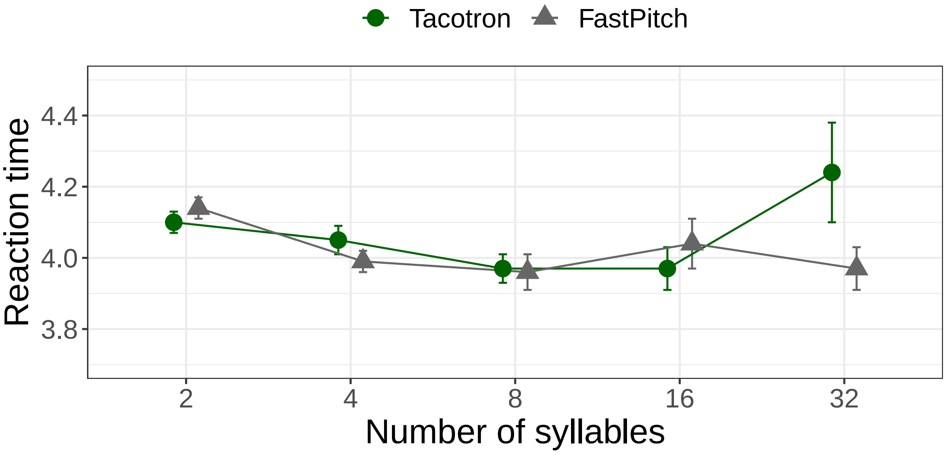


Figure 5.21: **Reaction times in tilt-deviant stimuli:** Comparative display across acoustic models with data pooled from participants in both the groups. Error bars represent absolute deviation from the median, normalized by number of trials per stimulus length.

In tilt-deviant utterances, increasing length supports faster decision making in both acoustic models. However, the relationship is not consistent, as longer utterances require longer processing time.

In Tacotron, the shortest 2-syllable stimuli are rated in a median of 4.14 seconds. Reaction time reduces incrementally, with stimulus length increasing up to 8-syllables. The minimum reaction time is 3.97 seconds and is sustained identically up to 16-syllables. As in Figure 5.21, we see a sharp rise of 270 ms as full-length utterances are presented to participants. However, diversity in responses for this group is also higher, as is reflected by a score of 0.14 on the median absolute deviation.

Next, in FastPitch, we also see that longer exposure results in faster decision making. Median reaction times reduce from 4.14 seconds in 2-syllable stimuli, to 3.96 in 8-syllables. They rise by 80 ms (with expected variance) in 16-syllables, but stabilize to 3.97 again for full-length utterances.

Confidence scores reflect very similar patterns as seen before for the sonority experiment. The proportion of participants who report "Very Confident" rises with length in both acoustic models. But the rise is sharper in FastPitch (38.2%) compared to Tacotron (20.9%).

Like seen in most experiments above, reaction times fall and confidence increases with increasing exposure to the stimuli. A sharp increase in Tacotron reaction times may be attributed to the high variance of the data, while that in 16-syllable FastPitch is more uniform, and needs to be investigated on its lexical content.

5.3.5.4 Influence of demographic variables

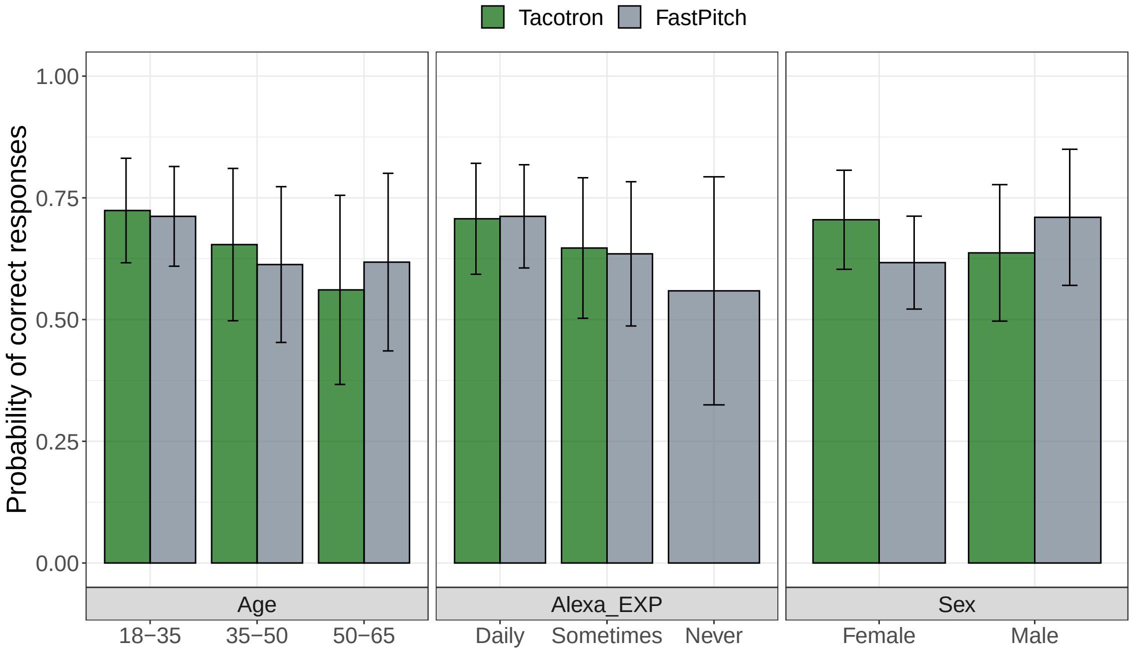


Figure 5.22: **Demographic variables in tilt-deviant stimuli:** Influence of age, exposure to TTS devices and sex of the participant on the likelihood of correct responses P_{ACC} in each acoustic model for the tilt-deviant WaveNet stimuli. Error bars represent the deviation of the P_{ACC} about standard error of the logistic regression model.

Figure 5.22 shows the unique influence of each demographic on the chance of providing correct responses. As seen in most cases before, older adults have a lower likelihood of scoring correct responses. They are 11.5% less likely to be accurate, and this difference is statistically significant in both acoustic models. Effects are considerably sharper in Tacotron [slope (SE) -.072 (0.19), p-val < 0.001], indicating that older people struggle more with tilt-deviant Tacotron stimuli. TTS exposure is only meaningful in the case of Group I participants, i.e. FastPitch stimuli, because all participants in Group reported some experience with TTS. No-exposure participants are 15.1%, less likely to provide correct answers. This is significantly lower [slope (SE) -0.67 (0.23), p-val < 0.01] than daily users of TTS devices. Occasional users also exhibit minimally significant lowering of P_{ACC}. Finally, sex-based differences are inconsistent across acoustic models. Male participants in FastPitch have a higher chance of scoring correctly, [slope (SE) 0.42 (0.14), p-val < 0.01], but lower in Tacotron [slope (SE) -0.31 (0.14), p-val < 0.05].

Therefore, older adults and no-exposure participants have a lower chance of scoring correctly. Age-based results are sharper in Tacotron, while exposure-based are meaningful only for FastPitch. On the other hand, sex of the participant is inconsistent in determining their accuracy in this experimental condition.

5.3.6 Tilt-alike stimuli

5.3.6.1 An overview of stimuli that do not deviate in spectral tilt

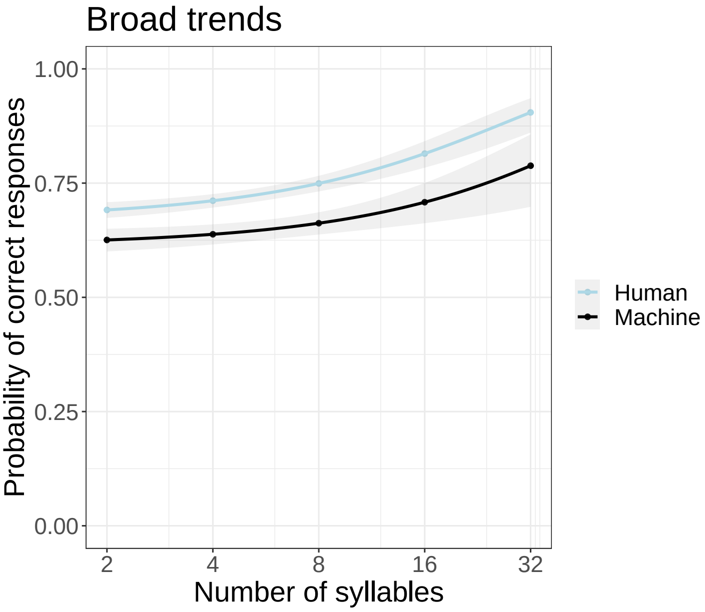


Figure 5.23: **Human vs Machine in tilt-alike stimuli:** A GLM-model fit comparing human and tilt-alike WaveNet stimuli, combined over both groups and acoustic models. Model fit describes the relationship between the predicted probability of correct responses (0=incorrect, 1=correct) and increasing stimulus-length. Generally, P_{ACC} can be seen rising for tilt-alike WaveNet stimuli.

This is a discussion on those stimuli which do not deviate from the human voice in terms of segmental spectral tilt. Figure 5.23 shows a rising trend, indicating that increasing length allows for more machine-likeness, and consequently a greater chance of accurate responses.

This relationship is also statistically significant [slope (SE) 0.03 (0.01), p-val < 0.01]. We see an increase of 16.25% between the endpoints of stimuli lengths, where P_{ACC} rises from 0.63 to 0.79. A comparison of adjacent stimuli lengths suggest incremental increase with every doubling of length. The sharpest increase is between 16 and 32, where the rise is 7.99%.

Participant data also suggests that most participants respond accurately. The proportion of participants who provide correct answers for at least half the trials is above 85% in all the stimuli lengths. This means that most participants can detect the machine-likeness of the stimuli.

Taking these results together, we can see that stimuli whose segmental spectral tilt is quite similar to the human voice, also appears quite machine-like. A careful observation of demographic variables and individual acoustic models is required now.

5.3.6.2 Comparison of Tactron and FastPitch stimuli

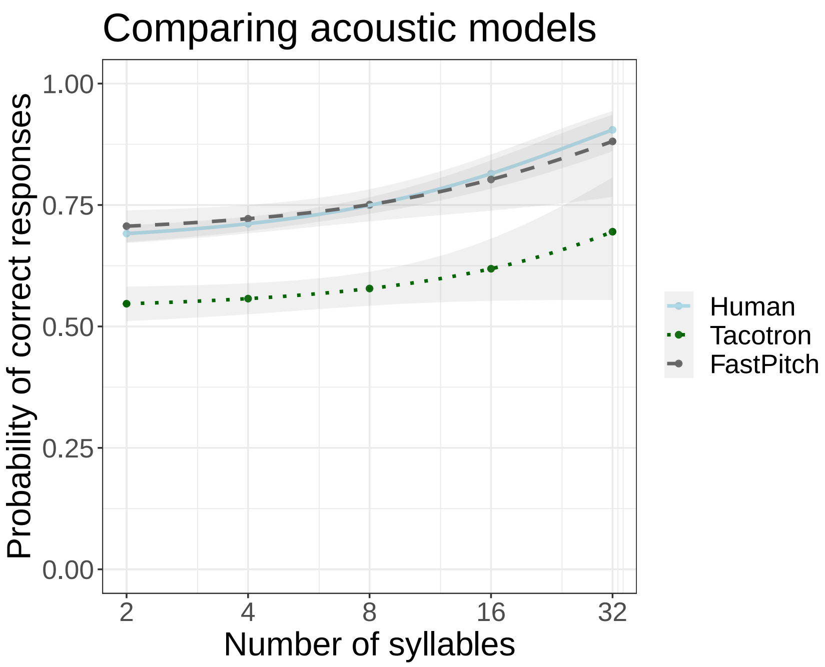


Figure 5.24: **Tacotron vs FastPitch in tilt-alike stimuli:** A GLM-model fit showing differences between the two acoustic models in the tilt-alike stimuli. P_{ACC} rises in both FastPitch and Tacotron, but reaches higher values in FastPitch.

Individual trends in Tacotron and FastPitch are uniform, but vary in magnitude. Figure 5.24 shows that progression in stimulus length increases the likelihood of obtaining correct responses in both acoustic models, tested over distinct groups of participants. However, the likelihood is visibly lower in Tacotron. We can also see that accuracy is predicted to be lower for shorter stimuli, compared to tilt-deviant Tacotron in the previous section.

In Tacotron, length is only a minimally significant influence (p-val < 0.1). The starting P_{ACC} is 0.54 for 2-syllable stimuli. It rises to 0.695, with a 14.82% rise between the endpoints of stimuli-lengths. Although P_{ACC} shows a rising pattern across adjacent stimuli lengths, there are no major peaks. The largest increase is simply between 16 and 32. However interestingly, participants who score above chance do not increase with length. This proportion is 64.5% at syllable-length 2, rises to 83.8% for 8-syllables, and then falls below 75% in full-length utterances. This indicates that longer tilt-similar utterances do appear human-like to some participants.

In FastPitch, stimulus length has a stronger influence on P_{ACC} [slope (SE), 0.04 (0.01), p-val < 0.05]. As noted before, P_{ACC} starts at 0.71, which is higher than the maximum predicted for Tacotron. It rises 17.45% to 0.88 for full-length utterances. Comparing adjacent lengths, we find that a high P_{ACC} of 0.75 is already seen at 8-syllables. The movement is however, gentle like above, with a large spike obtained in the expected 16 and 32 range. The curve is almost superimposed on the human-like utterances, indicating that machine-likeness is clearly perceptible. Similarly, participants who respond correctly on at least 50% of the trials is 96.7%, and remains above 90% for full-length ones.

These results show that tilt-similar stimuli do not appear human-like to our participants. However, a closer look at the demographic variables is necessary.

5.3.6.3 Reaction times and confidence scores

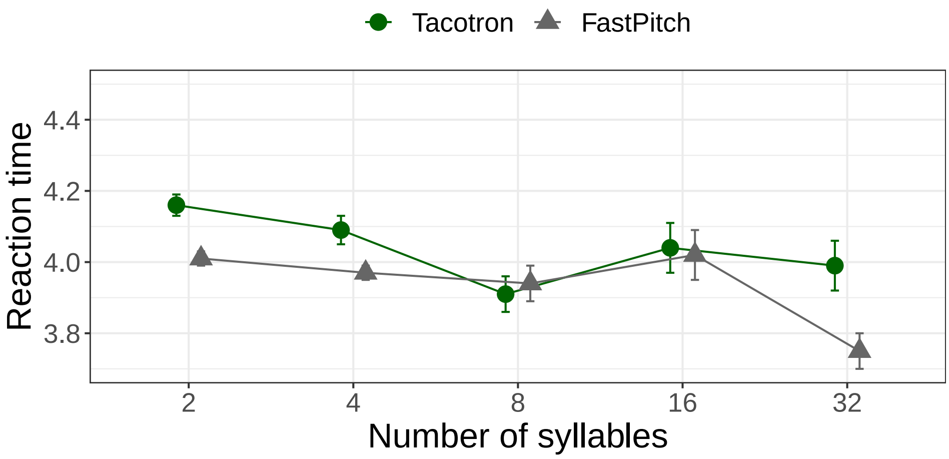


Figure 5.25: **Reaction times in tilt-alike stimuli:** Comparative display across acoustic models with data pooled from participants in both the groups. Error bars represent absolute deviation from the median, normalized by number of trials per stimulus length.

Reaction times for tilt-alike stimuli are uniform in both acoustic models. They fall up to middle-length, show a minimal increase at 16-syllables, and then drop again at 32-syllables. Overall, Tacotron is processed somewhat slower than FastPitch.

The shortest utterances require 4.16 seconds to be rated in Tacotron. Reaction time progressively declines to 3.91 seconds up to 8-syllable utterances. However, we see a notable rise of 130 ms at 16-syllable utterances, without much variance in the group. Full-length utterances are rated quickly, at 3.99 ms again.

Similarly, reaction times reduce up to 8-syllables, reaching a minimum of 3.94 seconds from 4.01. However, the increase at 16-syllables is somewhat gentler, i.e. of 80 ms. Variance is low overall, indicating uniform results. 32-syllable tilt-similar utterances are dismissed quickly as machine-like, reporting the lowest reaction score so far.

The self-reported confidence scores also support that Tacotron offers more confusion. The proportion of participants who report "Very Confident" is uniformly low, i.e. 38% and 30% in FastPitch and Tacotron respectively. While it only rises to 48.3% in Tacotron, it reaches 68.9% in FastPitch. Therefore, a majority of listeners report high confidence in rating FastPitch stimuli. As seen in Figure 5.25, reaction times are faster at all lengths of stimuli. At full-length utterances, participants are notably faster in FastPitch, i.e. by 240 ms in FastPitch compared to Tacotron. Another notable trend is that 16-syllable stimuli require the longest reaction times in both acoustic models. It is possible that the lexical content of a select set of utterances demand a longer processing times.

5.3.6.4 Influence of demographic variables

Figure 5.26 clearly exemplifies that all population groups are more accurate in detecting the machine-likeness of FastPitch utterances. Differences between groups are not as distinct as previously found.

Age-based differences are only significant in FastPitch [slope (SE) -0.91 (0.20), p-val < 0.001], where older adults are 18.8% less likely to respond correctly. Middle-aged adults also show modestly significant (p-val < 0.05) lowering of P_{ACC}, compared to those in 18-35 (i.e. young) age group. In Tacotron, neither of these results are significant. This indicates that young adults, an otherwise high-performing group, are also confused with Tacotron stimuli. Next, in terms of exposure to TTS devices, we see no significant differences in either of the acoustic models. Similarly, sex based differences are inconsistent across acoustic models. Only in Tacotron, do male participants have a 6.7% higher chance of responding correctly, with modestly significant effects [slope (SE) 0.27 (0.13), p-val < 0.05]. But in FastPitch, males show a non-significant but lower P_{ACC} by 3%. This shows that both male and female participants rate FastPitch stimuli with comparable accuracy, but Tacotron can present challenge for females.

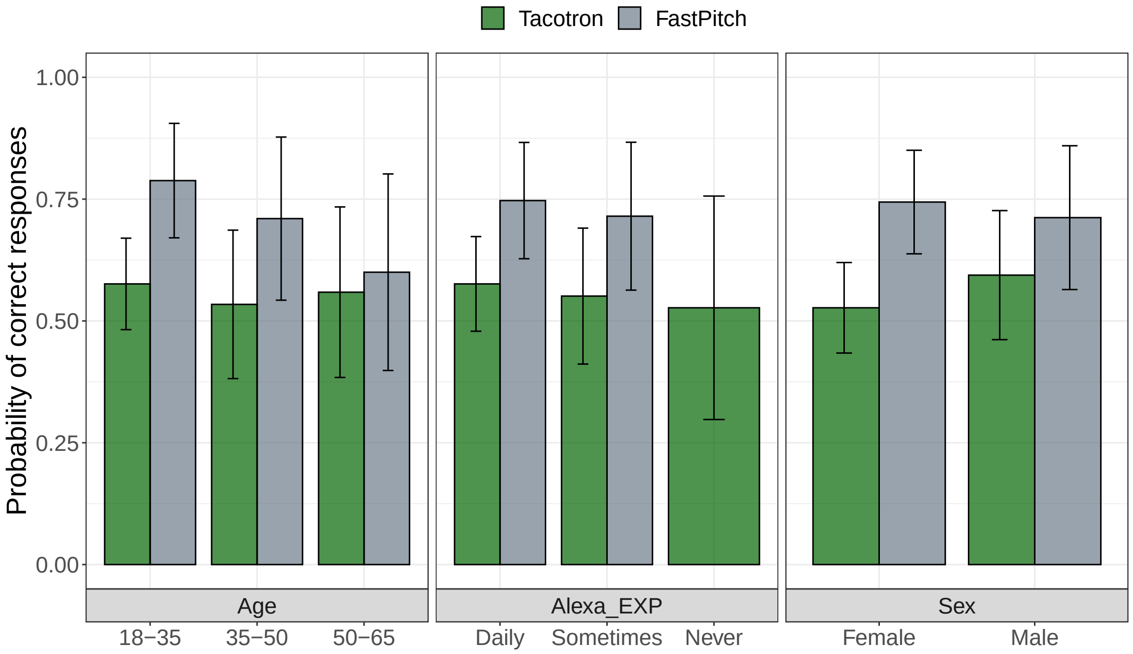


Figure 5.26: **Demographic variables in tilt-alike stimuli:** Influence of age, exposure to TTS devices and sex of the participant on the likelihood of correct responses P_{ACC} in each acoustic model for the tilt-alike WaveNet stimuli. Error bars represent the deviation of the P_{ACC} about standard error of the logistic regression model.

5.4 Discussion

5.4.1 Length and segmental distortion

Previous work on spoofing detection has demonstrated the realism of synthetic voices, and has been described as indistinguishable from human speech (Wang et al., 2020b) in many listening conditions (Terblanche et al., 2021). However, we found that machine-likeness could be detected in a majority of experimental conditions and groups. Out of the 6 phrase-sets tested, 4 corresponded with the increasing length hypothesis. Increasing stimulus length had a positive influence on the likelihood of participant accuracy in a majority of cases in WaveNet stimuli. In accordance with the Wecher-Febner's law, participants responded with minimal error to logarithmically increasing stimuli. Next, segmental distortion in terms of spectral tilt deviation and obstruent-richness only enhanced the difference between human and machine-likeness.

These results underscore the insufficient nature of traditional evaluation tests. Furthermore, these results also indicate that segmental distortion is perceivable in obstruent-rich and tilt-deviant utterances. It also supports previous research where distortion has pointed to higher-level attributes like naturalness and system-preferences (Bunnell et al., 1998).

The case of randomly selected phrases R2-P and the tilt-alike phrases is unclear. In the baseline condition, R2-P is generally rated more human-like with increasing stimulus length. Similarly, there are no major differences in the perception of tilt-deviant and tilt-alike stimuli. We investigated our stimuli presentation details to locate the potential influence of context, and quality of the preceding utterance. No sequential link between the previous utterances was observed to introduce contextual effects. Similarly, based on quality of the previous utterance, we could not locate any consistent ceiling effects. Since no obvious errors emerge from experiment design, it is possible that the stimuli contain artefacts that we did not account for. In previous work, spectral tilt deviation is computed per phrase, instead of the per segment average we chose. Deviation is incrementally introduced in the same phrase. Therefore, it will be important to recreate these experiments, with ideally a wider ranges of spectral tilt deviations. Finally, the data is imbalanced between long and short utterances. This may render our GLM model more sensitive to fluctuations in long utterances.

Phrase-set	Lexical content	Tacotron	FastPitch
R2-P	Within a short time she was walking briskly toward the Emerald City, her silver shoes tinkling merrily on the hard, yellow road bed.	23	16
R2-P	To what training she owed her skill.	17	17
R2-P	Was anything more than a tedious ridiculous nickname.	16	19
Tilt-Alike	Done up in gray wool until she resembled a small teddy bear	8	1
Tilt-Alike	He ceased to satirize himself because time dulled the irony of the situation and the joke lost its humor with its sting.	2	8
Tilt-Alike	his lordly patron's noble leg	3	8

Table 5.3: Expectation failed:Phrases that were rated very human-like in the baseline condition, or very machine-like in the tilt-alike condition across both groups and acoustic models. The columns on the right display the number of times each utterance was rated human-like.

5.4.2 Tacotron and FastPitch

Tacotron voices were judged to be human-like more frequently than FastPitch voices. This bears correspondence with the MOS evaluation in Table 4.1 where Tacotron WaveNet (Z) received a higher MOS rating compared to FastPitch WaveNet (Y). In a complementary study, we had shown that FastPitch selectively disrupts the micro-prosodic characteristics in voiceless fricatives. A potential explanation is that in FastPitch, an average F0 value is predicted for the entire duration of the phoneme, and is set to 0 in unvoiced regions. The interpolation must cause the raising at onset to be normalized with the steady-state regions of the vowel, as well as the offset. On the other hand, Tacotron2, with its auto-regressive nature predicts F0 based on previous samples, and retains the necessary microprosodic variation.

Non-autoregressive architectures like FastPitch are faster, and parallelizable, and offer to overcome several shortcomings in auto-regressive TTS (Łańcucki, 2021). For example, tighter constraints on phoneme length and duration (Ren et al., 2020) have contributed to previously reported problems like word repetitions or omissions (Ping et al., 2017). Moreover, greater control has been achieved in synthesizing expressive speech (Lee et al., 2021), especially in low-resource languages (Shah et al., 2021). The recourse is not limited to TTS, but is widespread in sequential data generation like machine translation and video captioning (Xiao et al., 2023; Yang et al., 2021). These rapid developments point to a compelling demand for parallelizable architectures. However, claims of parity between the two architectures must be revisited, based on our findings. Not only was FastPitch perceived more machine-like, it was also more sensitive to segmental distortion. To complement its development, we recommend that greater attention is paid to quality evaluation of non-autoregressive architectures.

5.4.3 Demographic variables

Older adults in our participant pool judged most stimuli to be human-like. Despite the imbalance in the dataset, the effect is consistent in most groups. These results are in line with declining sensitivity with age (Lin et al., 2013; Jerger et al., 1995; Huang and Tang, 2010). But a balanced study on age-related effects will inform many groups, especially those using AI devices for elderly support. We did not see consistent effects of prior exposure to TTS devices. As positive effects of exposure have been reported for tasks like word-recognition (Schwab et al., 1985), we had expected daily users to be the best performing group. However, we also observed that the effects of exposure are difficult to evaluate with self-report. More specific tests are needed.

Most conspicuously, we found a consistent effect of male listeners providing more accurate responses in human-machine detection. Since this trend was not limited to groups with a higher number of male participants younger age category, or those more exposed to TTS devices, we can eliminate a data-driven bias. Feminine features such as higher pitch and dispersion (Puts et al., 2011; Apicella and Feinberg, 2009) have been reported to appear more appealing to male participants. A possible explanation is that male participants could have higher engagement with the speaker of our experiments, and hence observed more fine-grained differences in its human-likeness. Other evidence comes from phonetics and hearing mechanisms which are particular to gender. Systematic studies on sex-based influences on hearing have shown differences in male and female hearing patterns (McFadden, 2014). Female participants have more sensitive hearing acuities (Chung et al., 1983; Stelmachowicz et al., 1989), that is slower to decline than their male counterparts (Pearson et al., 1995). Male participants on the other hand, report more robust perception in noise and sub-optimal listening conditions (Neff et al., 1996). Specific evidence from phonemic category perception identifies that male participants reduce VOT distinction between stop consonants, and consequently rely on F0 as a cue for voicing (Yu, 2022). This reliance is further reinforced, if the speaker is rated attractive, trustworthy or confident. In other words, male participants have greater experience of exploiting the F0 for contrastive information. Extrapolating from these studies, we speculate that male participants may be using the prosodic information more efficiently, to detect human-machine likeness in our stimuli. But more experiments are needed to investigate sex-based influences on the perception of synthetic speech. Care must be taken to closely monitor educational background, exposure to music and hearing sensitivity and other interacting factors that may bias their results.

5.4.4 Future implications

The perceptual significance of deviating obstruents in WaveNet systems has implications for multiple fields. First, it may motivate TTS engineers to focus on segmental attributes of a system, or even perform a post-processing of their audio. For example, (Fujimoto et al., 2018) demonstrate that the use of WaveNet vocoders with distinct periodic/aperiodic decomposition, scores higher naturalness. From a TTS evaluation perspective, the test methodology presented may offer a more fine-grained insight into localizing the source and perceptual significance of distortion, compared to traditional, MOS-based listening tests. Finally, if segmental characteristics of sonorants are indeed indistinguishable from human speech, then analysis of synthetically produced sonorants may generalize well to human speech. This could accelerate research in phonetics, because of the reduced reliance on speech data collection. It must be noted, however, that more variance can be seen in participant responses for synthetic speech. A potential reason is the imbalance between short and long utterances. This is a limitation of the dataset, as naturally occurring corpora do not contain utterances that are neatly balanced for obstruent/sonorant-richness, unless specially designed. In future work, it will be useful to redesign these experiments, with equal numbers of long and short stimuli, which are not “cut-outs" from running speech.

6 | Conclusion

6.1 Summary

In this thesis, we have studied segmental evaluation in Text-to-Speech synthesizers. Current practices in TTS evaluation are dominated by a listening test, where participants are requested to score the quality of speech on an ordered scale. This approach has received active criticism, because it does not provide any insight leading to system improvement. Our aim has been to develop a set of diagnostic frameworks that can identify specific weaknesses of TTS synthesizers. Our approaches incorporate perspectives from speech and behavioural sciences, for the frameworks of acoustic analysis and subjective evaluation respectively. These frameworks are particularly designed to evaluate the human-likeness of the voice generated by TTS synthesizers. The human and TTS voice data is provided by the original Blizzard 2013 challenge, and its recent extension to neural TTS. The human voice belongs to a female voice actor, who is a native speaker of American English. All TTS voices are generated with this voice as training data.

One of the first goals of this thesis was to disambiguate the concept of naturalness, which is a widely tested attribute in TTS synthesizers. A clear definition of this concept is seldom provided, and the response relies on the users' own interpretation of the term. We describe how naturalness is a multi-faceted perceptual attribute in TTS voices, and is driven equally by its appropriateness to various contexts. In targeted applications of TTS, naturalness is closely linked with human-likeness. For example, a full human-like functionality of the voice is required to support patients of various speech impairments. We identify that human-likeness is a desirable attribute in TTS synthesizers. Specifying such a goal further complemented the diagnostic nature of our proposed designs. First, because we could analyze the signal directly in comparison to the human voice. This means that we maintained the acoustic-phonetic features of the human voice as the constant, standard reference, and compared the same features in TTS voices against this reference. The deviation from the features of human voice became the comparative metric for evaluating TTS synthesizers. Then, in the proposed subjective evaluation, we could ask an unambiguous question Does this sound like a human or machine?

In Chapter 3, we introduce the Dive Into Divisions approach for segmental evaluation of TTS synthesizers. A segment is defined as a phoneme boundary, and segmentation of corpora is achieved through forced alignment. Obstruent consonants, especially stops and affricates, require a further step of sub-phonemic segmentation. After the boundaries are determined, a set of contrastive features are extracted from these segments of phonemes. Contrastive features are those acoustic-phonetic features which are responsible for meaningful differences in the categorical perception of speech sounds. For example, we hear the difference between /p/ and /b/, because of the difference in the duration of the voicing onset time. Similarly, formant features vary to characterise different vowel shapes. Contrastive features provide a characteristic representation of phonemes, and have been well-documented through several decades of phonetics research. Therefore, for every segment (vowels and obstruent consonants), we extract a set of contrastive features relevant to its category. Then, we compare them between the human voice and every TTS voice of the Blizzard Challenge 2013 dataset. Important diagnostic trends are revealed, which point to the TTS generation technique. For example, we find that vowels produced by HMM synthesizers are tightly clustered around their mean values. This can be traced back to the statistical averaging technique, where instances of vowels are produced using a limited set of model parameters in HMM synthesis. In other words, the acoustics inform us of the articulation mechanism, as in a classic phonetics project. Another important finding of this approach is that voiceless regions of obstruent consonants produced by WaveNet and WaveGAN vocoders, deviate strongly from the human voice. At this stage, a complete explanation is not clear. However, it can be confirmed that a feature-wise comparison between human and machine voices can lead to diagnostic insights about a global articulation failure, i.e, aperiodic regions in neural TTS.

The next challenge was to identify whether this distortion is perceivable to human listeners. None of the existing designs of subjective evaluation are suitable for testing distortion, because they depend on eliciting user response on complete utterances. The prosodic support, the contextual expectations and the repetitive nature of the task could distract the listener from identifying the segmental distortion in the signal. Therefore, we designed the Long Arms framework, where participants are presented with stimuli of varying lengths. In accordance with the Weber-Fechner law, which states that human perception is inherently logarithmic, our stimuli are also designed as logarithmic variants of stimulus lengths. This means that, stimuli of 2, 4, 8, 16 and 32 syllables are presented to our listeners in random order. Moreover, stimuli rich in obstruents (obstruent-rich) are compared with those poor in obstruents (sonorant-rich) to target sensitivity to obstruent distortion. Participants are asked: Does this sound like a human or machine? for every stimulus, and their responses are captured in a 2-AFC task. An analysis of these responses using a logistic regression model reveals indeed, that segmental distortion is perceivable as greater machine-likeness in WaveNet stimuli. This further underscores the importance of segmental analysis, and designs inspired from behavioural sciences for TTS evaluation.

6.1.1 Context

The original contributions in this thesis have been enabled by several key advancements in multiple disciplines. First, parallel data was available from the Blizzard 2013 Challenge in the human and TTS voices. This greatly reduced the speaker and lexical variation in the corpus, allowing us to focus on differences within the signal alone. It also provided a 300-hour training data, ensuring its extension neural TTS. This allowed our analytical work to be situated within the contemporary progress in TTS. Additionally, the cross-combination of autoregressive and non-autoregressive architectures provided several grounds for systematic feature comparison. Next, we draw inspiration from the active progress in the fields of corpus phonetics. There are several tools available for forced alignment, segmentation and feature analysis of large scale corpora. Their accompanying tutorials are also often provided. This enabled us to automate the acoustic analysis, and helped to overcome reliance on manual annotations. This helped greatly to bridge the gap between speech science and technology. Thirdly, an up-to-date and comprehensive compilation of obstruent features provided the groundwork for acoustic-phonetic analysis. We encourage future researchers of fundamental phonetics to visualize the usefulness of their findings in TTS evaluation. Finally, the most critical catalyst to this research is the growing faith in phonetics and speech science within TTS evaluation, and speech technology in general. This has enabled cross-disciplinary research in our own lab, and was further bolstered through a long collaboration with KTH, Sweden. This inclusion has been the most powerful contributor to work in this thesis.

6.1.2 Limitations

The present work is dependent on the availability of the parallel data, with the same lexical content available in the human and TTS voices. At this stage, we do not know the results of an investigation which does not have identical lexical content. Further complexities can arise if the source speaker does not match the target TTS generated voice, as is often the case in voice cloning, or low-resource adaptation of training data. At this stage, this thesis does not provide methods for speaker and content normalisation. This is a problem, because if deviation is observed in some acoustic-phonetic features, we cannot determine whether this is a distortion or a consequence of speaker differences. A second limitation appears with inadequate computational tools, especially for sub-phonemic boundary detection. As discussed in Section 3.2.3.2, the peak corresponding to the burst had to be estimated, by manually examining 20% of the obstruent spectrogram. This step involves intervention by a phonetics graduate, and is still sensitive to variations. Even after this, some instances were not properly demarcated, and had to be discarded. This is further exacerbated in postvocalic (VC) contexts. It is therefore very important that tools like AutoVOT, and DrVOT which automatically detect these sub-phonemic boundaries are rigorously developed for TTS evaluation. Finally, all of our stimuli for the perception tests were created using the existing utterances in the Blizzard 2013 test corpus. Therefore, a proper separation between obstruent-rich and sonorant-rich stimuli was constrained by their availability in the corpus. For example, both voiced and unvoiced obstruents were collapsed as a single class opposing sonorants, while the original finding of distortion was limited to voiceless regions only. Although we ascertained that the voiced obstruents were weighted lower than voiceless ones, a clearer separation would yield more robust results.

6.2 Future work

6.2.1 Scaling: segments, speakers and languages

The present work concerns itself with one female speaker. The analysis is limited to English, and only obstruent consonants and their neighbouring vowels are evaluated. Scaling this design to incorporate other segments, multiple speakers, and other languages is important to ascertain that the effects we observe are robust across several conditions. First, the articulation of nasal consonants depends on the individual shapes of sinus cavities. This means that nasal consonants and their contextual vowels retain speaker-sensitive information. Therefore, they hold particular relevance in applications of TTS like voice cloning, speaker anonymization and spoofing. In the present work, we used obstruent consonants to examine how TTS synthesizers handle very specific acoustic consequences, such as short term transience, aperiodicity etc, which were not possible to study in other phonological classes. Similarly, the nasals and liquids are produced through a coupling mechanism with the pharyngeal and the side chambers respectively. This mechanism introduces unique consequences like additional formants, antiresonances and wider bandwidths (Zsiga, 2013; Stevens, 2000). Therefore, evaluating these phonological classes through their acoustic-phonetic can aid in further diagnosis of neural TTS synthesizers.

6.2.2 Development as a no-reference metric

This thesis presents a feature-by-feature comparison of each TTS voice directly with the human voice. Drawing from Section 2.4.2.2, this approach bears resemblance with the full-reference metrics like PESQ, POLQA and ViSQOL. However, several TTS voices are created using pre-trained models where a direct comparison with the source speaker is not possible.

To overcome this issue, we suggest the following approach: create a model-based reference, and estimate deviation from the model parameters. A stepwise description is as follows:

- Download a large-scale, multi-speaker dataset such as LibriTTS, FoR
- Identify phonemic and sub-phonemic boundaries, and extract acoustic-phonetic features as described in Chapters 3 and 4.
- For every feature in every phoneme (e.g. spectral tilt of /f/) create a speaker independent probability distribution, but control for demographics like gender, accent and age.
- Examine the ranges and central tendencies of each feature, and evaluate them against existing literature. Remove aberrant speakers.
- Create a TTS corpus of comparable distribution. Segment and extract features, and cast them into a phoneme-wise distribution.
- Compare the reference, i.e, the speaker-independent human speech model with the equivalent feature model of the desired TTS voice. This comparison can be quantified using metrics like KL Divergence etc.

6.2.3 Using segmental distortion as a loss function

The present thesis is limited to an analysis of segments, but does not use that information to improve the TTS synthesizer. If distortion is expressed as a continuous, differentiable function then it can be introduced into the loss function of the TTS synthesizer. Then, reducing the KL divergence will optimize the output of the TTS, so that segmental distortion is also minimized.

6.2.4 Using LLMs as stimuli creators

As discussed before, we could not completely balance our stimuli to evaluate the effects of distortion of specific segments. Even naturally occurring, large corpora of English may not have a concentration of segments that we wish to evaluate. However, using Large Language Models like ChatGPT and Bard can alleviate some creative design. We requested ChatGPT to create a dialogue between a user and a voice based agent Alexa, such that Alexa's responses are full of sonorant consonants. Here was the response.

- **User:** Alexa, What's the latest news update?
- **Alexa:** Currently, there's **n**ote**w**o**r**thy **n**ews about a **n**ew tech**n**o**l**ogy, **r**evea**ling** **n**ew and **r**ema**r**kab**l**e adva**n**ce**m**e**n**ts.

Conversely, when asked for an obstruent-rich utterances, we received:

- **User**: Alexa, what’s the latest news update?
- **Alexa**: **P**re**s**en**t**ly, **th**ere’**s** **sig**ni**f**i**c**an**t** news **s**urroun**d**ing a **b**rea**kth**rough in **s**u**st**aina**b**le **tech**nology, **f**o**st**ering **th**e **f**uture of **k**ey indu**st**ries with **p**ioneering develo**p**men**ts**.

This approach can help us analyze segmental distortion in TTS within a conversational setting. Although it took a few attempts, the use of LLMs can augment the sentence generation, which would be quite laborious for a human writer.

6.2.5 Recipe for using segmental evaluation as an evaluation framework

The pipeline for segmental evaluation can be used for an acoustic as well as a perceptual evaluation of synthesized audio. A following set of steps are presented, where it can be used as a for-reference (ground-truth available) method:-

- Run the Dive into Divisions approach for an acoustic-phonetic feedback during the speech synthesis stage.
- Identify locations of distortion, and conduct a targeted improvement using knowledge based features.
- Re-evaluate the features of the synthesized voices. Locate distortions.
- State clear objectives for subjective evaluation, using specific use-cases for naturalness.
- Conduct subjective evaluation in multiple contextual and conversational settings.

6.3 Final remark

This PhD work shows how features of the small, segmental units of speech can be used to evaluate Text-to-Speech synthesizers. In particular, we focus on the human-likeness aspect of the multi-faceted naturalness. We discuss that human-likeness remains a desirable and important target of TTS synthesis, and caters to a range of diverse applications. Then, we describe the extraction and analysis feature which draws inspiration from standard techniques in phonetics. We identify the sites where non-neural and neural TTS synthesizers show statistically significant deviations from the human voice. Each TTS voice is compared directly with the human voice, which is always maintained as the reference. Then, we design a subjective evaluation framework which is inspired from the techniques in psychophysics, and is suitable to perceptually evaluate distorted segments. We show that segmental distortion is perceivable as increased machine-likeness even in modern, neural TTS. We hope that this

PhD work helps to recognize the contribution of phoneticians and cognitive scientists, and encourage them to conduct their research with TTS evaluation as another goal.

Bibliography

Alam, S., Johnston, B., Vitale, J., and Williams, M.-A. (2021). Would you trust a robot with your mental health? the interaction of emotion and logic in persuasive backfiring. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 384–391. IEEE.

Antons, J.-N., Schleicher, R., Arndt, S., Moller, S., Porbadnigk, A. K., and Curio, G. (2012). Analyzing speech quality perception using electroencephalography. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):721–731.

Apicella, C. L. and Feinberg, D. R. (2009). Voice pitch alters mate-choice-relevant perception in hunter–gatherers. *Proceedings of the Royal Society B: Biological Sciences*, 276(1659):1077–1082.

Arees, E. A. (1967). Investigation of speech transmission testing. Technical report, NORTHEASTERN UNIV BOSTON MASS.

Atal, B. S. and David, N. (1978). On finding the optimum excitation for lpc speech synthesis. *The Journal of the Acoustical Society of America*, 63(S1):S79–S79.

Awad, Z., Taghi, A. S., Sethukumar, P., Ziprin, P., Darzi, A., and Tolley, N. S. (2014). Binary versus 5-point likert scale in assessing otolaryngology trainees in endoscopic sinus surgery. *Otolaryngology—Head and Neck Surgery*, 151(1_suppl):P113–P113.

Aylett, M. P. and Yamagishi, J. (2008). Combining statistical parameteric speech synthesis and unit-selection for automatic voice cloning. *Proc. LangTech*, 2008.

Bachan, J. and Tokarski, M. (2017). Creation and evaluation of marytts speech synthesis for polish. In *Language and Technology Conference*.

Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummins, N., and Schuller, B. (2018). The perception and analysis of the likeability and human likeness of synthesized speech. In *Interspeech*.

Baljekar, P. and Black, A. W. (2016). Utterance selection techniques for tts systems using found speech. In *Speech Synthesis Workshop*, pages 184–189.

Basnet, M., Poudel, N., Dahal, S., and Subedi, S. (2023). Aawaj: Augmentative communication support for the vocally impaired using nepali text-to-speech.

Bayston, T. and Campanella, S. (1957). Continuous analysis speech band-width compression system. *The Journal of the Acoustical Society of America*, 29(11):1255–1256.

Beck, G. T. D., Wennberg, U., Malisz, Z., and Henter, G. E. (2022). Wavebender gan: An architecture for phonetically meaningful speech manipulation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6187–6191. IEEE.

Beerends, J. G., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J., and Keyhl, M. (2013). Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment. *Journal of the Audio Engineering Society*, 61(6):366–384.

Behrens, S. I., Egsvang, A. K. K., Hansen, M., and Møllegård-Schroll, A. M. (2018). Gendered robot voices and their influence on trust. In *Companion of the 2018 ACM/IEEE international conference on human-robot interaction*, pages 63–64.

Benoit, C. (1990). An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity. *Speech Communication*, 9(4):293–304.

Benoît, C., Grice, M., and Hazan, V. (1996). The sus test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech communication*, 18(4):381–392.

Betz, S., Carlmeyer, B., Wagner, P., and Wrede, B. (2018). Interactive hesitation synthesis: modelling and evaluation. *Multimodal Technologies and Interaction*, 2(1):9.

Beutnagel, M., Conkie, A., and Syrdal, A. K. (1998). Diphone synthesis using unit selection. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.

Biecek, P. (2018). Dalex: explainers for complex predictive models in r. *The Journal of Machine Learning Research*, 19(1):3245–3249.

Black, A. W. and Tokuda, K. (2005). The blizzard challenge-2005: Evaluating corpus-based speech synthesis on common datasets. In *Ninth European Conference on Speech Communication and Technology*.

Blumstein, S. E. and Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, 66(4):1001–1017.

Boersma, P. and Weenink, D. (2018). Praat: Doing phonetics by computer [computer program]. version 6.0. 37. *RetrievedFebruary*, 3:2018.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20(3):255.

Brunow, D. A. and Cullen, T. A. (2021). Effect of text-to-speech and human reader on listening comprehension for students with learning disabilities. *Computers in the Schools*, 38(3):214–231.

Bulhak, A. C. (1996). On the simulation of postmodernism and mental debility using recursive transition networks.

Bunnell, H. T., Hoskins, S. R., and Yarrington, D. (1998). Prosodic vs. segmental contributions to naturalness in a diphone synthesizer. In *ICSLP*.

Bush, A. M. (1972). Time-frequency resolution in speech analysis and synthesis. Technical report, Georgia Institute of Technology.

Camp, J., Kenter, T., Finkelstein, L., and Clark, R. (2023). Mos vs. ab: Evaluating text-to-speech systems reliably using clustered standard errors. In *Proc. Interspeech*, volume 2023.

Capes, T., Coles, P., Conkie, A., Golipour, L., Hadjitarkhani, A., Hu, Q., Huddleston, N., Hunt, M., Li, J., Neeracher, M., et al. (2017). Siri on-device deep learning-guided unit selection text-to-speech system. In *Interspeech*, pages 4011–4015.

Carlson, C. W. (1968). Computer use in parallel-formant speech synthesis. *The Journal of the Acoustical Society of America*, 44(1):391–391.

Cassel, L. E. and Steele, R. W. (1963). Evaluation of the effects of dynamic encoding on vocoder performance. *The Journal of the Acoustical Society of America*, 35(11):1911–1911.

Chalamandaris, A., Tsiakoulis, P., Karabetsos, S., Raptis, S., and LTD, I. (2013). The ilsp/innoetics text-to-speech system for the blizzard challenge 2013. In *Blizzard Challenge Workshop*. Citeseer.

Chandra, S. and Lin, W. (1977). Linear prediction with a variable analysis frame size. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(4):322–330.

Chang, R. C.-S., Lu, H.-P., and Yang, P. (2018). Stereotypes or golden rules? exploring likable voice traits of social robots as active aging companions for tech-savvy baby boomers in taiwan. *Computers in Human Behavior*, 84:194–210.

Chasaide, A. N., Chiaráin, N. N., Berthelsen, H., Wendler, C., and Murphy, A. (2015). Speech technology as documentation for endangered language preservation: The case of irish. In *ICPhS*, volume 2015, page 18th.

Chen, C.-n., Chu, C.-Y., Yeh, S.-L., Chu, H.-H., and Huang, P. (2012). Modeling the qoe of rate changes in skype/silk voip calls. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 119–128.

Chen, H.-C., Chen, C.-Y., Tsou, K.-M., and Chen, O.-C. (1997). A 0.75 kbps speech codec using recognition and synthesis schemes. In *1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings. Back to Basics: Attacking Fundamental Problems in Speech Coding*, pages 27–28. IEEE.

Chen, L., Evanini, K., and Sun, X. (2010). Assessment of non-native speech using vowel space characteristics. In *IEEE Spoken Language Technology Workshop*, pages 139–144. IEEE.

Chen, L.-H., Ling, Z., Jiang, Y., Song, Y., Xia, X.-J., Zu, Y.-Q., Yan, R.-Q., and Dai, L.-R. (2013). The ustc system for blizzard challenge 2013. In *Blizzard Challenge Workshop*.

Chen, Q. Q. and Park, H. J. (2021). How anthropomorphism affects trust in intelligent personal assistants. *Industrial Management & Data Systems*, 121(12):2722–2737.

Cho, T. and Ladefoged, P. (1999). Variation and universals in vot: evidence from 18 languages. *Journal of phonetics*, 27(2):207–229.

Chodroff, E. (2018). Corpus phonetics tutorial. *arXiv preprint arXiv:1811.05553*.

Chodroff, E. and Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in american english. *The Journal of the Acoustical Society of America*, 136(5):2762–2772.

Chung, D. Y., Mason, K., Gannon, R. P., and Willson, G. N. (1983). The ear effect as a function of age and hearing loss. *The Journal of the Acoustical Society of America*, 73(4):1277–1282.

Clark, R., Silen, H., Kenter, T., and Leith, R. (2019). Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs. pages 99–104.

Cohen, M. F., Mickunas Jr, J., Miller, J., and Voiers, W. D. (1965). Diagnostic rhyme test for the evaluation of communications systems. *The Journal of the Acoustical Society of America*, 37(6):1206–1206.

Cohn, M. and Zellou, G. (2020). Perception of concatenative vs. neural text-to-speech (tts): Differences in intelligibility in noise and language attitudes. In *International Conference on Speech Communication and Technology (Interspeech)*, pages 1733–1737.

Cole, J. and Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology*, 7(1).

Combescure, P., Le Guyader, A., and Gilloire, A. (1982). Quality evaluation of 32 kbit/s coded speech by means of degradation category ratings. In *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 988–991. IEEE.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., and Tang, Y. (2013). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):572–585.

Cooper, E., Huang, W.-C., Toda, T., and Yamagishi, J. (2022). Generalization ability of mos prediction networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8442–8446. IEEE.

Cooper, E., Huang, W.-C., Tsao, Y., Wang, H.-M., Toda, T., and Yamagishi, J. (2023). The voicemos challenge 2023: Zero-shot subjective speech quality prediction for multiple domains. *Proceedings of InterSpeech: The VoiceMOS Challenge*.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *The Journal of the Acoustical Society of America*, 24(6):597–606.

Dall, R., Yamagishi, J., and King, S. (2014). Rating naturalness in speech synthesis: The effect of style and expectation. In *Proceedings of Speech Prosody*. Citeseer.

Danielescu, A. (2020). Eschewing gender stereotypes in voice assistants to promote inclusion. In *Proceedings of the 2nd conference on conversational user interfaces*, pages 1–3.

David, E., Schroeder, M., Logan, B., and Prestigiacomo, A. (1962). Voice-excited vocoders for practical speech bandwidth reduction. *IRE Transactions on Information Theory*, 8(5):101–105.

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., and Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3):331.

Debnath, A., Patil, S. S., Nadiger, G., and Ganesan, R. A. (2020). Low-resource end-to-end sanskrit tts using tacotron2, waveglow and transfer learning. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–5. IEEE.

Dehaene, S. (2003). The neural basis of the weber–fechner law: a logarithmic mental number line. *Trends in cognitive sciences*, 7(4):145–147.

Delattre, P., Cooper, F. S., Liberman, A. M., and Gerstman, L. (1954). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, 26(1):137–137.

Dettweiler, H. and Hess, W. (1985). Concatenation rules for demisyllable speech synthesis. *Acta Acustica united with Acustica*, 57(4-5):268–283.

Dewey, G. (1923). *Relativ frequency of English speech sounds*. Harvard University Press.

DiCanio, C., Nam, H., Whalen, D. H., Bunnell, H. T., Amith, J. D., and García, R. C. (2013). Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America*, 134(3):2235–2246.

Ditz, H. M. and Nieder, A. (2016). Numerosity representations in crows obey the weber–fechner law. *Proceedings of the Royal Society B: Biological Sciences*, 283(1827):20160083.

Donovan, R. E. (1996). *Trainable speech synthesis*. PhD thesis, Citeseer.

Donovan, R. E. and Woodland, P. C. (1995). Improvements in an hmm-based speech synthesiser. In *Fourth European Conference on Speech Communication and Technology*.

Dorman, M. F., Soli, S., Dankowski, K., Smith, L. M., McCandless, G., and Parkin, J. (1990). Acoustic cues for consonant identification by patients who use the ineraid cochlear implant. *The Journal of the Acoustical Society of America*, 88(5):2074–2079.

Drugman, T. and Raitio, T. (2014). Excitation modeling for hmm-based speech synthesis: breaking down the impact of periodic and aperiodic components. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 260–264. IEEE.

Dudley, H., Riesz, R. R., and Watkins, S. S. (1939). A synthetic speaker. *Journal of the Franklin Institute*, 227(6):739–764.

Duffy, S. A. and Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, 35(4):351–389.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope*.

Epley, N., Waytz, A., and Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864.

Escudero, P., Boersma, P., Rauber, A. S., and Bion, R. A. (2009). A cross-dialect acoustic description of vowels: Brazilian and european portuguese. *The Journal of the Acoustical Society of America*, 126(3):1379–1393.

Etzrodt, K. and Engesser, S. (2021). Voice-based agents as personified things: Assimilation and accommodation as equilibration of doubt. *Human-Machine Communication*, 2:57–76.

Evers, V., Reetz, H., and Lahiri, A. (1998). Crosslinguistic acoustic categorization of sibilants independent of phonological status. *Journal of phonetics*, 26(4):345–370.

Fairbanks, G. (1958). Test of phonemic differentiation: The rhyme test. *The Journal of the Acoustical Society of America*, 30(7):596–600.

Fairbanks, G. and Kodman Jr, F. (1957). Word intelligibility as a function of time compression. *The Journal of the acoustical Society of America*, 29(5):636–641.

Fant, G. (1953). Speech communication research. *Royal Swedish Academy of Engineering Sciences*, 2:331–337.

Fechner, G. T. (1860). *Elemente der psychophysik*, volume 2. Breitkopf u. Härtel.

Fernández-Torné, A. and Matamala, A. (2015). Text-to-speech vs. human voiced audio descriptions: a reception study in films dubbed into catalan. *The Journal of Specialised Translation*, 24:61–88.

Flanagan, J. (1960). A resonance-vocoder and baseband complement: A hybrid systems for speech transmission. *IRE Transactions on Audio*, (3):95–102.

Fletcher, H. and Steinberg, J. (1929). Articulation testing methods. *The Bell System Technical Journal*, 8(4):806–854.

Fron, C. and Korn, O. (2019). A short history of the perception of robots and automata from antiquity to modern times. *Social robots: technological, societal and ethical aspects of human-robot interaction*, pages 1–12.

Fujimoto, T., Yoshimura, T., Hashimoto, K., Oura, K., Nankaku, Y., and Tokuda, K. (2018). Speech synthesis using wavenet vocoder based on periodic/aperiodic decomposition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 644–648. IEEE.

Fussell, J., Abzug, B., Boudra, P., and Cowing, M. (1978). Providing channel error protection for a 2400 bps linear predictive coded voice system. In *ICASSP'78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 462–465. IEEE.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93.

Gessinger, I., Raveh, E., O'Mahony, J., Steiner, I., and Möbius, B. (2016). A shadowing experiment with natural and synthetic stimuli. *Phonetik & Phonologie*, 12:58–61.

Gessinger, I., Raveh, E., Steiner, I., and Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing. *Speech Communication*, 127:43–63.

Giannouli, V. and Banou, M. (2020). The intelligibility and comprehension of synthetic versus natural speech in dyslexic students. *Disability and Rehabilitation: Assistive Technology*, 15(8):898–907.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Govender, A. and King, S. (2018). Using pupillometry to measure the cognitive load of synthetic speech. In *International Conference on Speech Communication and Technology (Interspeech)*, pages 2838–2842.

Govender, A., Wagner, A. E., and King, S. (2019). Using pupil dilation to measure cognitive load when listening to text-to-speech in quiet and in noise. *International Conference on Speech Communication and Technology (Interspeech)*, pages 1551–1555.

Gracco, V. L. (1994). Some organizational characteristics of speech movement control. *Journal of Speech, Language, and Hearing Research*, 37(1):4–27.

Graham, Y., Awad, G., and Smeaton, A. (2018). Evaluation of automatic video captioning using direct assessment. *PloS one*, 13(9):e0202789.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2014). Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451.

Greene, B. G., Logan, J. S., and Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, & Computers*, 18(2):100–107.

Greenspan, S. L., Bennett, R. W., and Syrdal, A. K. (1989). A study of two standard speech intelligibility measures. *The Journal of the Acoustical Society of America*, 85(S1):S43–S43.

Greenspan, S. L., Bennett, R. W., and Syrdal, A. K. (1998). An evaluation of the diagnostic rhyme test. *International Journal of Speech Technology*, 2(3):201–214.

Greenspan, S. L., Nusbaum, H. C., and Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):421.

Grether, C. and Stroh, R. (1973). Subjective evaluation of differential pulse-code modulation using the speech" goodness" rating scale. *IEEE transactions on audio and electroacoustics*, 21(3):179–184.

Grice, M. (1989). Syntactic structures and lexicon requirements for semantically unpredictable sentences in a number of languages. In *Speech Input/Output Assessment and Speech Databases*.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243.

Griffiths, J. D. (1966). Further rhyme-test modification for diagnostic articulation testing. *The Journal of the Acoustical Society of America*, 40(5):1256–1256.

Griffiths, J. D. (1968). Optimum linear filter for speech transmission. *The Journal of the Acoustical Society of America*, 43(1):81–86.

Gunderson, J. (1991). Limits of intelligibility of accelerated synthesized speech by inexperienced sighted and experienced blind listeners. In *Proceedings of the Human Factors Society Annual Meeting*, volume 35, pages 496–500. SAGE Publications Sage CA: Los Angeles, CA.

Guo, H., Soong, F. K., He, L., and Xie, L. (2019). Exploiting syntactic features in a parsed tree to improve end-to-end tts. *ArXiv*, abs/1904.04764.

Gupta, R., Arndt, S., Antons, J.-N., Schleicher, R., Möller, S., Falk, T. H., et al. (2013). Neurophysiological experimental facility for quality of experience (qoE) assessment. In *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*, pages 1300–1305. IEEE.

Gupta, R., Banville, H. J., and Falk, T. H. (2015). Physyqx: A database for physiological evaluation of synthesised speech quality-of-experience. In *2015 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, pages 1–5. IEEE.

Gutierrez, E., Gallegos, P. O., and Lai, C. (2021). Location, location: Enhancing the evaluation of text-to-speech synthesis using the rapid prosody transcription paradigm. *ArXiv*, abs/2107.02527.

Hain, T., Woodland, P. C., Evermann, G., Gales, M. J., Liu, X., Moore, G. L., Povey, D., and Wang, L. (2005). Automatic transcription of conversational telephone speech. *IEEE Transactions on Speech and Audio Processing*, 13(6):1173–1185.

Handley, Z. (2009). Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication*, 51(10):906–919.

Hanley, C. N. (1956). Factorial analysis of speech perception. *journal of Speech and Hearing Disorders*, 21(1):76–87.

Harris, C. M. (1953). A study of the building blocks in speech. *The Journal of the Acoustical Society of America*, 25(5):962–969.

Hauptmann, A. G. (1993). Speakez: A first experiment in concatenation synthesis from a large corpus. In *Third European Conference on Speech Communication and Technology*.

Hayashi, T., Watanabe, S., Toda, T., Takeda, K., Toshniwal, S., and Livescu, K. (2019). Pretrained text embeddings for enhanced text-to-speech synthesis. In *INTERSPEECH*, pages 4430–4434.

Hazan, V. and Grice, M. (1989). The assessment of synthetic speech intelligibility using semantically unpredictable sentences. In *Speech Input/Output Assessment and Speech Databases*.

Helms, S. W. (1968a). Conferencing with pitch-excited channel vocoders. *Journal of the Audio Engineering Society*, 16(3):296–300.

Helms, S. W. (1968b). Effects of speech combining within analysis-synthesis processes. *The Journal of the Acoustical Society of America*, 44(1):391–391.

Hines, A., Skoglund, J., Kokaram, A. C., and Harte, N. (2015). Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):1–18.

Hinterleitner, F. (2017). *Quality of Synthetic Speech: Perceptual Dimensions, Influencing Factors, and Instrumental Assessment*. Springer.

Hinterleitner, F., Möller, S., Falk, T. H., and Polzehl, T. (2010). Comparison of approaches for instrumentally predicting the quality of text-to-speech systems: Data from blizzard challenges 2008 and 2009. In *Blizzard Challenge Workshop*, volume 2010, pages 48–60.

Hinterleitner, F., Norrenbrock, C., and Möller, S. (2013). Is intelligibility still the main problem? a review of perceptual quality dimensions of synthetic speech. In *Proc. 8th ISCA Workshop on Speech Synthesis (SSW 8)*, pages 147–151.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., and Benson, R. W. (1952). Development of materials for speech audiometry. *Journal of speech and hearing disorders*, 17(3):321–337.

Holmes, J. (1973). The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. *IEEE transactions on Audio and Electroacoustics*, 21(3):298–305.

House, A. S., Williams, C. E., Hecker, M. H., and Kryter, K. D. (1965). Articulation-testing methods: consonantal differentiation with a closed-response set. *The Journal of the Acoustical Society of America*, 37(1):158–166.

Howard, C. G. (1956). Speech analysis-synthesis scheme using continuous parameters. *The Journal of the Acoustical Society of America*, 28(6):1091–1098.

Huang, Q. and Tang, J. (2010). Age-related hearing loss or presbycusis. *European Archives of Oto-rhino-laryngology*, 267:1179–1191.

Huang, W.-C., Cooper, E., Tsao, Y., Wang, H.-M., Toda, T., and Yamagishi, J. (2022). The voicemos challenge 2022. *Proceedings of InterSpeech: The VoiceMOS Challenge*.

Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 373–376. IEEE.

Im, H., Sung, B., Lee, G., and Kok, K. Q. X. (2023). Let voice assistants sound like a machine: Voice and task type effects on perceived fluency, competence, and consumer attitude. *Computers in Human Behavior*, 145:107791.

ITU, I. (1994). A method for subjective performance assessment of the quality of speech voice output devices. *International Telecommunication Union Std*.

Iverson, P., Smith, C. A., and Evans, B. G. (2006). Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration. *The Journal of the Acoustical Society of America*, 120(6):3998–4006.

Iwahashi, N., Kaiki, N., and Sagisaka, Y. (1992). Concatenative speech synthesis by minimum distortion criteria. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 2, pages 65–68. IEEE Computer Society.

Jansen, D. (2019). Discovering the uncanny valley for the sound of a voice. *Unpublished master's thesis]. School of Humanities and Digital Sciences Department of Cognitive Science & Artificial Intelligence. Tilburg.*

Jekosch, U. (1992). The cluster-identification test. In *Second International Conference on Spoken Language Processing*.

Jenkins, J. J. and Franklin, L. D. (1982). Recall of passages of synthetic speech. *Bulletin of the Psychonomic Society*, 20(4):203–206.

Jensen, U., Moore, R. K., Dalsgaard, P., and Lindberg, B. (1993). Modelling of intonation contours at the sentence level using CHMMs and the 1961 o'connor and arnold scheme. In *Proc. 3rd European Conference on Speech Communication and Technology (Eurospeech 1993)*, pages 785–788.

Jensen, U., Moore, R. K., Dalsgaard, P., and Lindberg, B. (1994). Modelling intonation contours at the phrase level using continuous density hidden markov models. *Computer Speech & Language*, 8(3):247–260.

Jerger, J., Chmiel, R., Wilson, N., and Luchi, R. (1995). Hearing impairment in older adults: new concepts. *Journal of the American Geriatrics Society*, 43(8):928–935.

Jongman, A. (1989). Duration of frication noise required for identification of english fricatives. *The Journal of the Acoustical Society of America*, 85(4):1718–1725.

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263.

Jreige, C., Patel, R., and Bunnell, H. T. (2009). Vocalid: Personalizing text-to-speech synthesis for individuals with severe speech impairment. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pages 259–260.

Kain, A. and van Santen, J. P. (2007). Unit-selection text-to-speech synthesis using an asynchronous interpolation model. In *SSW*, pages 172–177. Citeseer.

Karhila, R., Remes, U., and Kurimo, M. (2013). Noise in hmm-based speech synthesis adaptation: Analysis, evaluation methods and experiments. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):285–295.

Kawahara, H. (2006). Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353.

Keating, D., Evans, A., Wyper, D., and Cunningham, E. (1986). A comparison of the intelligibility of some low cost speech synthesis devices. *British journal of disorders of communication*, 21(2):167–172.

Keeler, L., Clement, G., Strong, W., and Palmer, E. (1976). Two preliminary studies of the intelligibility of predictor-coefficient and formant-coded speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):429–432.

Keeler, L. O., Strong, W. J., and Palmer, E. P. (1974). Comparison of the intelligibility of predictor coefficient and formant coded speech. *The Journal of the Acoustical Society of America*, 56(S1):S15–S15.

Kim, D.-S. (2005). Anique: An auditory model for single-ended speech quality estimation. *IEEE Transactions on Speech and Audio Processing*, 13(5):821–831.

Kim, H., Martin, K., Hasegawa-Johnson, M., and Perlman, A. (2010). Frequency of consonant articulation errors in dysarthric speech. *Clinical linguistics & phonetics*, 24(10):759–770.

King, S. (2010). A beginners' guide to statistical parametric speech synthesis. *The Centre for Speech Technology Research, University of Edinburgh, UK*.

King, S. (2014). Measuring a decade of progress in text-to-speech. *Loquens*, 1(1):e006–e006.

King, S. and Karaiskos, V. (2013). The blizzard challenge 2013. In *The Blizzard Challenge Workshop*. `http://festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf`.

King, S., Tokuda, K., Zen, H., and Yamagishi, J. (2008). Unsupervised adaptation for hmm-based speech synthesis. ISCA.

Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., and Lee, K. A. (2017). The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection.

Kirby, J. P. and Ladd, D. R. (2016). Effects of obstruent voicing on vowel f 0: Evidence from "true voicing" languages. *The Journal of the Acoustical Society of America*, 140(4):2400–2411.

Kirkland, A., Lameris, H., Székely, E., and Gustafson, J. (2022). Where's the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence. In *Proceedings of Interspeech*, pages 18–22.

Kishore, S., Black, A. W., Kumar, R., and Sangal, R. (2003). Experiments with unit selection speech databases for indian languages.

Klatt, D. H. (1987). Review of text-to-speech conversion for english. *The Journal of the Acoustical Society of America*, 82(3):737–793.

Klatt, D. H., Tiao, J., and Tetschner, W. (1984). Using dectalk as an aid for the handicapped. *The Journal of the Acoustical Society of America*, 75(S1):S85–S85.

Koenig, L. L., Shadle, C. H., Preston, J. L., and Mooshammer, C. R. (2013). Toward improved spectral measures of/s: Results from adolescents.

Koizumi, Y., Zen, H., Yatabe, K., Chen, N., and Bacchiani, M. (2022). SpecGrad: Diffusion Probabilistic Model based Neural Vocoder with Adaptive Noise Spectral Shaping. In *Proc. Interspeech 2022*, pages 803–807.

Kong, J., Kim, J., and Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. (2021). Diffwave: A versatile diffusion model for audio synthesis. *International Conference on Learning Representations*.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Kryter, K. D. (1956). On predicting the intelligibility of speech from acoustical measures. *Journal of Speech and Hearing Disorders*, 21(2):208–217.

Kuhl, P. K. (1993). Innate predispositions and the effects of experience in speech perception: The native language magnet theory. *Developmental neurocognition: Speech and face processing in the first year of life*, pages 259–274.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental science*, 9(2):F13–F21.

Kühne, K., Fischer, M. H., and Zhou, Y. (2020). The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in neurorobotics*, 14:105.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Ladefoged, P. and Maddieson, I. (1996). The sounds of the world's languages.

Lameris, H., Mehta, S., Henter, G. E., Gustafson, J., and Székely, É. (2023). Prosody-controllable spontaneous tts with neural hmms. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lawrence, W. (1953). The synthesis of speech from signal which have a low information rate. *Communication theory*, pages 460–469.

Le Maguer, S., King, S., and Harte, N. (2022). Back to the Future: Extending the Blizzard Challenge 2013. In *Proc. Interspeech 2022*, pages 2378–2382.

Le Maguer, S., King, S., and Harte, N. (2024). The limits of the mean opinion score for speech synthesis evaluation. *Computer Speech & Language*, 84:101577.

Lee, E. J., Nass, C., and Brave, S. (2000). Can computer-generated speech have gender? an experimental test of gender stereotype. In *CHI'00 extended abstracts on Human factors in computing systems*, pages 289–290.

Lee, K., Park, K., and Kim, D. (2021). Styler: Style factor modeling with rapidity and robustness via speech decomposition for expressive and controllable neural text to speech. *arXiv preprint arXiv:2103.09474*.

Lehiste, I. and Peterson, G. E. (1959a). Linguistic considerations in the study of speech intelligibility. *The Journal of the Acoustical Society of America*, 31(3):280–286.

Lehiste, I. and Peterson, G. E. (1959b). Vowel amplitude and phonemic stress in american english. *The Journal of the Acoustical Society of America*, 31(4):428–435.

Li, F. and Allen, J. B. (2011). Manipulation of consonants in natural speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):496–504.

Li, N. and Loizou, P. C. (2008). The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise. *The Journal of the Acoustical Society of America*, 124(6):3947–3958.

Li, N. and Loizou, P. C. (2010). Masking release and the contribution of obstruent consonants on speech recognition in noise by cochlear implant users. *The Journal of the Acoustical Society of America*, 128(3):1262–1271.

Li, W., Lei, S., Huang, Q., Zhou, Y., Wu, Z., Kang, S., and Meng, H. M. (2023). Towards spontaneous style modeling with semi-supervised pre-training for conversational text-to-speech synthesis. *InterSpeech*, abs/2308.16593.

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8):1.

Liberman, M. Y. (2019). Corpus phonetics. *Annual Review of Linguistics*, 5:91–107.

Lin, F. R., Yaffe, K., Xia, J., Xue, Q.-L., Harris, T. B., Purchase-Helzner, E., Satterfield, S., Ayonayon, H. N., Ferrucci, L., Simonsick, E. M., et al. (2013). Hearing loss and cognitive decline in older adults. *JAMA internal medicine*, 173(4):293–299.

Lindblom, B. and Maddieson, I. (1988). Phonetic universals in consonant systems. *Language, speech and mind*, 6278.

Ling, Z.-H. and Wang, R.-H. (2006). Hmm-based unit selection using frame sized speech segments. In *Ninth international conference on spoken language processing*.

Ling, Z.-H. and Wang, R.-H. (2007). Hmm-based hierarchical unit selection combining kullback-leibler divergence with likelihood criterion. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1245. IEEE.

Liu, H.-M., Tsao, F.-M., and Kuhl, P. K. (2005). The effect of reduced vowel working space on speech intelligibility in mandarin-speaking young adults with cerebral palsy. *The Journal of the Acoustical Society of America*, 117(6):3879–3889.

Liu, R., Sisman, B., Bao, F., Gao, G., and Li, H. (2020). Modeling prosodic phrasing with multi-task learning in tacotron-based tts. *IEEE Signal Processing Letters*, 27:1470–1474.

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., and Wang, H.-M. (2019). MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion. In *Proc. Interspeech 2019*, pages 1541–1545.

Logan, J. S., Greene, B. G., and Pisoni, D. B. (1989). Segmental intelligibility of synthetic speech produced by rule. *The Journal of the Acoustical Society of America*, 86(2):566–581.

Lu, Y. and Cooke, M. (2009). The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, 51(12):1253–1262.

Lux, F., Koch, J., and Vu, N. T. (2022). Low-resource multilingual and zero-shot multispeaker TTS. In He, Y., Ji, H., Li, S., Liu, Y., and Chang, C.-H., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 741–751, Online only. Association for Computational Linguistics.

Magee, M., Lewis, C., Noffs, G., Reece, H., Chan, J. C., Zaga, C. J., Paynter, C., Birchall, O., Rojas Azocar, S., Ediriweera, A., et al. (2020). Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. *The Journal of the Acoustical Society of America*, 148(6):3562–3568.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580.

Maki, H., Sakti, S., Tanaka, H., and Nakamura, S. (2018). Quality prediction of synthesized speech based on tensor structured EEG signals. *PloS one*, 13(6).

Malfait, L., Berger, J., and Kastner, M. (2006). P. 563—the itu-t standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1924–1934.

Malisz, Z., Henter, G. E., Valentini-Botinhao, C., Watts, O., Beskow, J., and Gustafson, J. (2019). Modern speech synthesis for phonetic sciences: A discussion and an evaluation. In *International Congress of Phonetic Sciencesnces ICPhS 2019 5-9 August 2019, Melbourne, Australia Melbourne Convention and Exhibition Centre*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Mao, Y., Tian, S., Qin, Y., and Han, J. (2019). A new sensory sweetness definition and sweetness conversion method of five natural sugars, based on the weber-fechner law. *Food chemistry*, 281:78–84.

Mara, M., Schreibelmayr, S., and Berger, F. (2020). Hearing a nose? user expectations of robot appearance induced by different robot voices. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 355–356.

Markel, J. D. and Gray, A. J. (1976). *Linear prediction of speech*, volume 12. Springer Science & Business Media.

Markopoulos, K., Maniati, G., Vamvoukakis, G., Ellinas, N., Vardaxoglou, G., Kakoulidis, P., Oh, J., Jho, G., Hwang, I., Chalamandaris, A., Tsiakoulis, P., and Raptis, S. (2023). Generating Multilingual Gender-Ambiguous Text-to-Speech Voices. In *Proc. INTERSPEECH 2023*, pages 621–625.

Mattheyses, W. and Verhelst, W. (2015). Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217.

Mattys, S. L., Carroll, L. M., Li, C. K., and Chan, S. L. (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. *Speech communication*, 52(11-12):887–899.

Mattys, S. L. and Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of memory and Language*, 65(2):145–160.

Mayo, C., Clark, R. A., and King, S. (2005). Multidimensional scaling of listener responses to synthetic speech.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *International Conference on Speech Communication and Technology (Interspeech)*, pages 498–502.

McCarthy, D. T. P. D. (2019). *The acoustics of place of articulation in English plosives.* PhD thesis, Newcastle University.

McFadden, D. (2014). Sex differences in the auditory system. In *Gonadal Hormones and Sex Differences in Behavior*, pages 261–298. Psychology Press.

McGee, V. E. (1964). Semantic components of the quality of processed speech. *Journal of Speech and Hearing Research*, 7(4):310–323.

McGinn, C. and Torre, I. (2019). Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots. In *2019 14th ACM/IEEE international Conference on human-robot interaction (HRI)*, pages 211–221. IEEE.

McGonegal, C., Rabiner, L., and Rosenberg, A. (1977). A subjective evaluation of pitch detection methods using lpc synthesized speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):221–229.

Melnik-Leroy, G. A. and Navickas, G. (2023). Can better perception become a disadvantage? synthetic speech perception in congenitally blind users. *INTERSPEECH 2023*.

Ménard, L., Trudeau-Fisette, P., Côté, D., and Turgeon, C. (2016). Speaking clearly for the blind: Acoustic and articulatory correlates of speaking conditions in sighted and congenitally blind speakers. *PLoS One*, 11(9):e0160088.

Mendelson, J. and Aylett, M. P. (2017). Beyond the listening test: An interactive approach to TTS evaluation. In *International Conference on Speech Communication and Technology (Interspeech)*, pages 249–253.

Mille, S., Belz, A., Bohnet, B., Graham, Y., Pitler, E., and Wanner, L. (2018). The first multilingual surface realisation shared task (sr'18): Overview and evaluation results. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12.

Mitchell, W. J., Szerszen Sr, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., and MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception*, 2(1):10–12.

Moore, R. K. and Skidmore, L. (2019). On the Use/Misuse of the Term 'Phoneme'. In *Proc. Interspeech 2019*, pages 2340–2344.

Morise, M., Yokomori, F., and Ozawa, K. (2016). World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884.

Moss, H. B., Aggarwal, V., Prateek, N., González, J., and Barra-Chicote, R. (2020). Boffin tts: Few-shot speaker adaptation by bayesian optimization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7639–7643.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6):453–467.

Moulines, E. and Verhelst, W. (1995). Time-domain and frequency-domain techniques for prosodic modification of speech. *Speech coding and synthesis*, pages 519–555.

Munson, W. and Karlin, J. (1962). Isopreference method for evaluating speech-transmission circuits. *The Journal of the Acoustical Society of America*, 34(6):762–774.

Nass, C., Moon, Y., and Green, N. (1997). Are machines gender neutral? gender-stereotypic responses to computers with voices. *Journal of applied social psychology*, 27(10):864–876.

Nearey, T. M. and Shammass, S. E. (1987). Formant transitions as partly distinctive invariant properties in the identification of voiced stops. *Canadian Acoustics*, 15(4):17–24.

Neff, D. L., Kessler, C. J., and Dethlefs, T. M. (1996). Sex differences in simultaneous masking with random-frequency maskers. *The Journal of the Acoustical Society of America*, 100(4):2547–2550.

Nguyen, C. M., Phung, L. V., Bui, C. T., Truong, T. V., and Nguyen, H. T. (2021). Learning vietnamese-english code-switching speech synthesis model under limited code-switched data scenario. In *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II 18*, pages 153–163. Springer.

Nickerson, J. F. and MILLER Jr, A. (1960). A comparison of five articulation tests. Technical report, MONTANA STATE COLL BOZEMAN.

Noah, B., Sethumadhavan, A., Lovejoy, J., and Mondello, D. (2021). Public perceptions towards synthetic voice technology. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 65, pages 1448–1452. SAGE Publications Sage CA: Los Angeles, CA.

Nusbaum, H., Schwab, E., and Pisoni, D. (1984). Subjective evaluation of synthetic speech: Measuring preference, naturalness and acceptibility. In *Proceedings of the Institution of Electrical Engineers*, volume 10, pages 391–407. Speech Research Lab-Tech, Bloomington IN.

Nusbaum, H. C., Francis, A. L., and Henly, A. S. (1997). Measuring the naturalness of synthetic speech. *International journal of speech technology*, 2(1):7–19.

Nusbaum, H. C. and Pisoni, D. B. (1985). Constraints on the perception of synthetic speech generated by rule. *Behavior Research Methods, Instruments, & Computers*, 17(2):235–242.

Nye, P. and Gaitenby, J. (1973). Consonant intelligibility in synthetic speech and in a natural speech control (modified rhyme test results). *Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications*, 101(134):77.

Okubo, T., Mochizuki, R., and Kobayashi, T. (2006). Hybrid voice conversion of unit selection and generation using prosody dependent hmm. *IEICE transactions on information and systems*, 89(11):2775–2782.

Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., et al. (2018). Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR.

O'Mahony, J., Oplustil-Gallegos, P., Lai, C., and King, S. (2021). Factors affecting the evaluation of synthetic speech in context. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 148–153.

P.10, I.-T. R. (2006). Vocabulary for performance and quality of service.

Pachl, W., Urbanek, G., and Rothauser, E. (1971). Preference evaluation of a large set of vocoded speech signals. *IEEE Transactions on Audio and Electroacoustics*, 19(3):216–224.

Pachl, W. P., Rothauser, E. H., and Urbanek, G. F. (1968). Comparison of preference scales. *The Journal of the Acoustical Society of America*, 44(1):385–385.

Painter, J., Gupta, S., and Wilson, L. (1973). Multipath modeling for aeronautical communications. *IEEE Transactions on Communications*, 21(5):658–662.

Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Pandey, A., Gogoi, P., and Tang, K. (2020). Understanding forced alignment errors in hindi-english code-mixed speech–a feature analysis. In *Proceedings of First Workshop on Speech Technologies for Code-Switching in Multilingual Communities 2020*, pages 13–17.

Pandey, A., Le Maguer, S., Carson-Berndsen, J., and Harte, N. (2021). Mind your p's and k's–comparing obstruents across tts voices of the blizzard challenge 2013. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 166–171.

Pandey, A., Le Maguer, S., Carson-Berndsen, J., and Harte, N. (2022). Production characteristics of obstruents in WaveNet and older TTS systems. In *Proc. Interspeech 2022*, pages 2373–2377.

Pandey, A., Le Maguer, S., Edlund, J., and Harte, N. (2023). Natural choice: Comparing place classification between natural and tacotron fricatives. *Proceedings of the 20th International Congress of Phonetic Sciences*, 20:3161–3165.

Parmonangan, I. H., Tanaka, H., Sakti, S., Takamichi, S., and Nakamura, S. (2019). Speech quality evaluation of synthesized japanese speech using EEG. *International Conference on Speech Communication and Technology (Interspeech)*, pages 1228–1232.

Patil, H. A., Patel, T. B., Shah, N. J., Sailor, H. B., Krishnan, R., Kasthuri, G., Nagarajan, T., Christina, L., Kumar, N., Raghavendra, V., et al. (2013). A syllable-based framework for unit selection synthesis in 13 indian languages. In *2013 International Conference Oriental CO-COSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–8. IEEE.

Patton, B., Agiomyrgiannakis, Y., Terry, M., Wilson, K. W., Saurous, R. A., and Sculley, D. (2016). AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech. *Proceedings of NeurIPS End-to-end Learning for Speech and Audio Processing Workshop,*, abs/1611.09207.

Pearson, J. D., Morrell, C. H., Gordon-Salant, S., Brant, L. J., Metter, E. J., Klein, L. L., and Fozard, J. L. (1995). Gender differences in a longitudinal study of age-associated hearing loss. *The Journal of the Acoustical Society of America*, 97(2):1196–1205.

Peirce, J., Hirst, R., and MacAskill, M. (2022). *Building experiments in PsychoPy*. Sage.

Perrotin, O., Stephenson, B., Gerber, S., and Bailly, G. (2023). The Blizzard Challenge 2023. In *Proc. 18th Blizzard Challenge Workshop*, pages 1–27.

Peterson, G. E. and Subrahmanyam, D. (1959). Evaluation of time-frequency scanning for narrow-band speech transmission. *The Journal of the Acoustical Society of America*, 31(1):113–113.

Peterson, G. E., Wang, W. S.-Y., and Sivertsen, E. (1958). Segmentation techniques in speech synthesis. *The Journal of the Acoustical Society of America*, 30(8):739–742.

Phung, T.-N., Phan, T.-S., Vu, T. T., Luong, M. C., and Akagi, M. (2013). Improving naturalness of hmm-based tts trained with limited data by temporal decomposition. *IEICE TRANSACTIONS on Information and Systems*, 96(11):2417–2426.

Picart, B., Drugman, T., and Dutoit, T. (2012). Assessing the intelligibility and quality of hmm-based speech synthesis with a variable degree of articulation. In *The Listening Talker*.

Ping, W., Peng, K., and Chen, J. (2018). Clarinet: Parallel wave generation in end-to-end text-to-speech. *ArXiv*, abs/1807.07281.

Ping, W., Peng, K., Gibiansky, A., Arik, S. Ö., Kannan, A., Narang, S., Raiman, J., and Miller, J. (2017). Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv: Sound*.

Pisoni, D. and Hunnicutt, S. (1980). Perceptual evaluation of mitalk: The mit unrestricted text-to-speech system. In *ICASSP'80. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 572–575. IEEE.

Pisoni, D. B., Nusbaum, H. C., and Greene, B. G. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, 73(11):1665–1676.

Plumpe, M., Acero, A., Hon, H.-W., and Huang, X. (1998). Hmm-based smoothing for concatenative speech synthesis. In *Fifth International Conference on Spoken Language Processing*.

Polkosky, M. D. and Lewis, J. R. (2003). Expanding the mos: Development and psychometric evaluation of the mos-r and mos-x. *International Journal of Speech Technology*, 6(2):161–182.

Pols, L. and Boxelaar, G. (1986). Comparative evaluation of the speech quality of speech coders and text-to-speech synthesizers. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 901–904. IEEE.

Pols, L. C. and Partners, S. (1992). Multi-lingual synthesis evaluation methods. In *Second International Conference on Spoken Language Processing*.

Porbadnigk, A. K., Antons, J.-N., Blankertz, B., Treder, M. S., Schleicher, R., Möller, S., and Curio, G. (2010). Using erps for assessing the (sub) conscious perception of noise. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 2690–2693. IEEE.

Potter, R. K. (1945). Visible patterns of sound. *Science*, 102(2654):463–470.

Pradhan, A., Findlater, L., and Lazar, A. (2019). " phantom friend" or" just a box with information" personification and ontological categorization of smart speaker-based voice assistants by older adults. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Prajwal, K. and Jawahar, C. (2021). Data-efficient training strategies for neural tts systems. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, pages 223–227.

Puts, D. A., Barndt, J. L., Welling, L. L., Dawood, K., and Burriss, R. P. (2011). Intrasexual competition among women: Vocal femininity affects perceptions of attractiveness and flirtatiousness. *Personality and Individual Differences*, 50(1):111–115.

Pérez Zarazaga, P., Malisz, Z., Henter, G. E., and Juvela, L. (2023). Speaker-independent neural formant synthesis. In *Proc. INTERSPEECH 2023*, pages 5556–5560.

Quintas, S. and Trancoso, I. (2020). Evaluation of deep learning approaches to text-to-speech systems for european portuguese. In *PROPOR*, pages 34–42.

Raptis, S., Chalamandaris, A., Tsiakoulis, P., and Karabetsos, S. (2012). The ilsp text-to-speech system for the blizzard challenge 2012. In *Blizzard Challenge Workshop*.

Raptis, S., Tsiakoulis, P., Chalamandaris, A., and Karabetsos, S. (2016). Expressive speech synthesis for storytelling: the innoetics'entry to the blizzard challenge 2016. In *Proc. Blizzard Challenge*.

Rec, I. (1996). P. 800: Methods for subjective determination of transmission quality. *International Telecommunication Union, Geneva*, 22.

Recasens, D., Pallarès, M. D., and Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. *The Journal of the Acoustical Society of America*, 102(1):544–561.

Redmon, C. (2020). Lexical acoustics: Linking phonetic systems to the higher-order units they encode. *PhD dissertation, University of Kansas, Lawrence*.

Reichl, P., Egger, S., Schatz, R., and D'Alconzo, A. (2010). The logarithmic nature of qoe and the role of the weber-fechner law in qoe assessment. In *2010 IEEE International Conference on Communications*, pages 1–5. IEEE.

Reimao, R. and Tzerpos, V. (2019). For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–10. IEEE.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2020). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech.

Repp, B. H. (1984). Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. *Language and speech*, 27(3):245–254.

Richards, D. (1964). Transmission performance assessment for telephone network planning. In *Proceedings of the Institution of Electrical Engineers*, volume 111, pages 931–940. IET.

Richards, D. (1974). Calculation of opinion scores for telephone connections. In *Proceedings of the Institution of Electrical Engineers*, volume 121, pages 313–323. IET.

Richards, D. and Buck, G. (1960). Telephone echo tests. *Proceedings of the IEE-Part B: Electronic and Communication Engineering*, 107(36):553–556.

Richards, D. and Swaffield, J. (1959). Assessment of speech communication links. *Proceedings of the IEE-Part B: Radio and Electronic Engineering*, 106(26):77–89.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

Ross, K. and Ostendorf, M. (1994). A dynamical system model for generating f0 for synthesis. In *The Second ESCA/IEEE Workshop on Speech Synthesis*.

Rothauser, E. (1969). Ieee recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17:225–246.

Rothauser, E., Urbanek, G., and Pachl, W. (1971). A comparison of preference measurement methods. *The Journal of the Acoustical Society of America*, 49(4B):1297–1308.

Rusko, M. et al. (2016). Development of the slovak hmm-based tts system and evaluation of voices in respect to the used vocoding techniques. *Computing and Informatics*, 35(6):1467–1490.

Sakti, S. and Nakamura, S. (2013). Towards language preservation: Design and collection of graphemically balanced and parallel speech corpora of indonesian ethnic languages. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–5. IEEE.

Sambur, M. and Jayant, N. (1976). Lpc analysis/synthesis from speech inputs containing quantizing noise or additive white noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(6):488–494.

Sanders, W. R., Benbassat, G. V., and Smith, R. L. (1976). Speech synthesis for computer assisted instruction: The miss system and its applications. *ACM SIGCUE Outlook*, 10(SI):200–211.

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., and Brenier, J. (2007). An acoustic study of real and imagined foreigner-directed speech. *Journal of the Acoustical Society of America*, 121(5):3044.

Scherer, S., Lucas, G. M., Gratch, J., Rizzo, A. S., and Morency, L.-P. (2015a). Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, 7(1):59–73.

Scherer, S., Morency, L.-P., Gratch, J., and Pestian, J. (2015b). Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4789–4793. IEEE.

Schreibelmayr, S. and Mara, M. (2022). Robot voices in our daily lives: Vocal human-likeness and application context as determinants of user acceptance. *Frontiers in Psychology*, page 1843.

Schroeder, J. and Epley, N. (2016). Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General*, 145(11):1427.

Schwab, E. C., Nusbaum, H. C., and Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human factors*, 27(4):395–408.

Seaborn, K., Miyake, N. P., Pennefather, P., and Otake-Matsuura, M. (2021). Voice in human–agent interaction: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–43.

Shadle, C. H. and Mair, S. J. (1996). Quantifying spectral characteristics of fricatives. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1521–1524. IEEE.

Shah, R., Pokora, K., Ezzerg, A., Klimkov, V., Huybrechts, G., Putrycz, B., Korzekwa, D., and Merritt, T. (2021). Non-autoregressive tts with explicit duration modelling for low-resource highly expressive speech. *arXiv preprint arXiv:2106.12896*.

Shamei, A., Liu, Y., and Gick, B. (2023). Reduction of vowel space in alzheimer's disease. *JASA Express Letters*, 3(3).

Sharma, B. and Prasanna, S. M. (2017). Enhancement of spectral tilt in synthesized speech. *IEEE Signal Processing Letters*, 24(4):382–386.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Sherwood, B. A. (1979). Computers: The computer speaks: Rapid speech synthesis from printed text input could accommodate an unlimited vocabulary. *IEEE spectrum*, 16(8):18–25.

Šimko, J., Törö, T., , Vainio, M., and Suni, A. (2020a). Prosody under control: A method for controlling prosodic characteristics in text-to-speech synthesis by adjustments in latent reference space. In *Proceedings of 10th International Conference on Speech Prosody 2020, Tokyo, Japan.* ISCA.

Šimko, J., Vainio, M., Suni, A., et al. (2020b). Analysis of speech prosody using wavenet embeddings: The lombard effect. In *Proceedings of 10th International Conference on Speech Prosody 2020, Tokyo, Japan*. ISCA.

Sonderegger, M. and Keshet, J. (2012). Automatic measurement of voice onset time using discriminative structured prediction. *The Journal of the Acoustical Society of America*, 132(6):3965–3979.

Stelmachowicz, P. G., Beauchaine, K. A., Kalberer, A., and Jesteadt, W. (1989). Normative thresholds in the 8-to 20-khz range as a function of age. *The Journal of the Acoustical Society of America*, 86(4):1384–1391.

Stevens, K. N. (2000). *Acoustic phonetics*, volume 30. MIT press.

Stevens, K. N. and Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, 64(5):1358–1368.

Stewart, J. Q. (1922). An electrical analogue of the vocal organs. *Nature*, 110(2757):311–312.

Streeter, L. A. (1988). Applying speech synthesis to user interfaces. In *Handbook of human-computer interaction*, pages 321–343. Elsevier.

Strong, W. J. (1967). Machine-aided formant determination for speech synthesis. *The Journal of the Acoustical Society of America*, 41(6):1434–1442.

Stylianou, Y. (1998). Concatenative speech synthesis using a harmonic plus noise model. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on speech and audio processing*, 9(1):21–29.

Subrahmanyam, D. and Peterson, G. (1959). Time-frequency scanning in narrow-band speech transmission. *IRE Transactions on Audio*, (6):148–160.

Suen, C. Y. and Beddoes, M. P. (1973). Development of a digital spelled-speech reading machine for the blind. *IEEE Transactions on Biomedical Engineering*, (6):452–459.

Sussman, H. M., Fruchter, D., and Cable, A. (1995). Locus equations derived from compensatory articulation. *The Journal of the Acoustical Society of America*, 97(5):3112–3124.

Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America*, 90(3):1309–1325.

Sydeserff, H., Caley, R., Isard, S. D., Jack, M. A., Monaghan, A. I., and Verhoeven, J. (1992). Evaluation of speech synthesis techniques in a comprehension task. *Speech communication*, 11(2-3):189–194.

Székely, É., Henter, G. E., Beskow, J., and Gustafson, J. (2019). Spontaneous conversational speech synthesis from found data. In *Interspeech*, pages 4435–4439.

Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A., and Broadbent, E. (2011). The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics*, 3:253–262.

Tan, X. (2023). *Neural Text-to-Speech Synthesis*. Springer Nature.

Tang, Y. (2021). Glimpse-based estimation of speech intelligibility from speech-in-noise using artificial neural networks. *Computer Speech & Language*, 69:101220.

Tanga, K. and Bennettb, R. (2019). UNITE AND CONQUER: BOOTSTRAPPING FORCED ALIGNMENT TOOLS FOR CLOSELY-RELATED MINORITY LANGUAGES (MAYAN). In *International Congress of Phonetic Sciences (ICPhS)*, pages 3584–3552.

Tay, B., Jung, Y., and Park, T. (2014). When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior*, 38:75–84.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

Terblanche, C., Harrison, P., and Gully, A. J. (2021). Human spoofing detection performance on degraded speech. In *Interspeech*, pages 1738–1742.

Terzić, K. and Hansard, M. (2016). Methods for reducing visual discomfort in stereoscopic 3d: A review. *Signal Processing: Image Communication*, 47:402–416.

Thorndike, E. L. (1932). A teacher's word book of the twenty thousand words found most frequently and widely in general reading for children and young people.

Tiomkin, S., Malah, D., Shechtman, S., and Kons, Z. (2011). A hybrid text-to-speech system that combines concatenative and statistical synthesis units. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:1278–1288.

Tits, N., El Haddad, K., and Dutoit, T. (2019). Exploring transfer learning for low resource emotional tts. In *Proceedings of SAI Intelligent Systems Conference*, pages 52–60. Springer.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1315–1318. IEEE.

Torre, I. and Le Maguer, S. (2020). Should robots have accents? In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 208–214. IEEE.

Trovato, G., Ramos, J., Azevedo, H., Moroni, A., Magossi, S., Ishii, H., Simmons, R., and Takanishi, A. (2015). Designing a receptionist robot: Effect of voice and appearance on anthropomorphism. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 235–240. IEEE.

Tryfou, G., Pellin, M., and Omologo, M. (2014). Time-frequency reassigned cepstral coefficients for phone-level speech segmentation. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 2060–2064. IEEE.

Tseng, W.-C., yu Huang, C., Kao, W.-T., Lin, Y. Y., and yi Lee, H. (2021). Utilizing Self-Supervised Representations for MOS Prediction. In *Proc. Interspeech 2021*, pages 2781–2785.

Turner, G. S., Tjaden, K., and Weismer, G. (1995). The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, 38(5):1001–1013.

Vainio, M., Järvikivi, J., Werner, S., Volk, N., and Välikangas, J. (2002). Effect of prosodic naturalness on segmental acceptability in synthetic speech. In *IEEE Workshop on Speech Synthesis*, pages 143–146. Citeseer.

Valentini-Botinhao, C., Yamagishi, J., King, S., and Maia, R. (2014). Intelligibility enhancement of hmm-generated speech in additive noise by modifying mel cepstral coefficients to increase the glimpse proportion. *Computer Speech & Language*, 28(2):665–686.

Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *CoRR*, abs/1609.03499.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. In *SSW*.

van Heuven, V. J. and van Bezooijen, R. (1995). Quality evaluation of synthesized speech. In *Speech coding and synthesis*, number 21, page 707738. Elsevier Amsterdam.

VaroSanec-SkariC, G. (1999). Relation between voice pleasantness and distribution of the spectral energy. In *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS 99), San Francisco.*

Varshney, L. R. and Sun, J. Z. (2013). Why do we perceive logarithmically? *Significance*, 10(1):28–31.

Venkatagiri, H. (1994). Effect of sentence length and exposure on the intelligibility of synthesized speech. *Augmentative and Alternative Communication*, 10(2):96–104.

Voiers, W. (1968). The present state of digital vocoding technique: A diagnostic evaluation. *IEEE Transactions on Audio and Electroacoustics*, 16(2):275–279.

Voiers, W. D. (1967). Performance evaluation of speech processing devices. iii. diagnostic evaluation of speech intelligibility. Technical report, SPERRY RAND RESEARCH CENTER SUDBURY MA.

Voiers, W. D., Cohen, M. F., and Mickunas, J. (1965). Evaluation of speech processing devices. 1. intelligibility, quality, speaker recognizability. Technical report, SPERRY RAND RESEARCH CENTER SUDBURY MA.

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Henter, G. E., Le Maguer, S., Malisz, Z., Székely, É., Tånnander, C., et al. (2019). Speech Synthesis Evaluation—State-of-the-Art Assessment and Suggestion for a Novel Research Program. In *Speech Synthesis Workshop (SSW)*.

Wallbridge, S., Bell, P., and Lai, C. (2021). It's Not What You Said, it's How You Said it: Discriminative Perception of Speech as a Multichannel Communication System. In *Proc. Interspeech 2021*, pages 2386–2390.

Wan, V., Latorre, J., Yanagisawa, K., Braunschweiler, N., Chen, L., Gales, M. J., and Akamine, M. (2013). Building hmm-tts voices on diverse data. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):296–306.

Wang, J., Yang, H., Shao, R., Abdullah, S., and Sundar, S. S. (2020a). Alexa as coach: Leveraging smart speakers to build social agents that reduce public speaking anxiety. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13.

Wang, L., Kim, S., and Zhou, X. (2023). Money in a "safe" place: Money anthropomorphism increases saving behavior. *International Journal of Research in Marketing*, 40(1):88–108.

Wang, X., Takaki, S., and Yamagishi, J. (2019). Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:402–415.

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K. A., et al. (2020b). Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. (2017). Tacotron: Towards end-to-end speech synthesis. In *INTERSPEECH*.

Watts, O., Stan, A., Clark, R., Mamiya, Y., Giurgiu, M., Yamagishi, J., and King, S. (2013). Unsupervised and lightly-supervised learning for rapid construction of tts systems in multiple languages fromfound'data: evaluation and analysis. In *8th ISCA Speech Synthesis Workshop: Barcelona, Spain*, pages 101–106. ISCA-INST SPEECH COMMUNICATION ASSOC.

Waytz, A., Heafner, J., and Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology*, 52:113–117.

Weidmüller, L. (2022). Human, hybrid, or machine?: Exploring the trustworthiness of voice-based assistants. *Human-Machine Communication*, 4:85–110.

Wilhelms-Tricarico, R., Reichenbach, J., and Marple, G. (2013). The lessac technologies hybrid concatenated system for blizzard challenge 2013. In *Blizzard Challenge Workshop*. Citeseer.

Williams, G. and Moye, L. (1971). Subjective evaluation of unsuppressed echo in simulated long-delay telephone communications. In *Proceedings of the Institution of Electrical Engineers*, volume 118, pages 401–408. IET.

Wilson, S. and Moore, R. K. (2017). Robot, alien and cartoon voices: implications for speech-enabled systems. In *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*, pages 40–44.

Wishna, S. (1973). Intelligibility improvement of analog communication systems using an amplitude control technique. *IEEE Transactions on Communications*, 21(5):655–658.

Wolf, J., Klovstad, J., Schwartz, R., Cosell, L., and Makhoul, J. (1978). Speech compression and synthesis. Technical report, BOLT BERANEK AND NEWMAN INC CAMBRIDGE MA.

Wong, D. and Markel, J. (1978). An intelligibility evaluation of several linear prediction vocoder modifications. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(5):424–435.

Wright, C., Altom, M. J., and Olive, J. (1986). Diagnostic evaluation of a synthesizer's acoustic inventory. *The Journal of the Acoustical Society of America*, 79(S1):S25–S25.

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., and Sizov, A. (2015). Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Xiao, Y., Wu, L., Guo, J., Li, J., Zhang, M., Qin, T., and Liu, T.-y. (2023). A survey on non-autoregressive generation for neural machine translation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Xu, J., Tan, X., Ren, Y., Qin, T., Li, J., Zhao, S., and Liu, T.-Y. (2020). Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2802–2812.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009a). Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):66–83.

Yamagishi, J., Nose, T., Zen, H., Toda, T., and Tokuda, K. (2008). Performance evaluation of the speaker-independent hmm-based speech synthesis system “hts 2007” for the blizzard challenge 2007. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3957–3960. IEEE.

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Guan, Y., Hu, R., Oura, K., Wu, Y.-J., et al. (2010). Thousands of voices for hmm-based speech synthesis–analysis and application of tts systems built on various asr corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):984–1004.

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Hu, R., Guan, Y., Oura, K., Tokuda, K., et al. (2009b). Thousands of voices for hmm-based speech synthesis. In *Tenth Annual Conference of the International Speech Communication Association.*

Yamamoto, R., Song, E., and Kim, J.-M. (2020). Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE.

Yan, Y., Tan, X., Li, B., Zhang, G., Qin, T., Zhao, S., Shen, Y., Zhang, W., and Liu, T.-Y. (2021). Adaptive text to speech for spontaneous style. In *Interspeech.*

Yanagisawa, K., Latorre, J., Wan, V., Gales, M. J., and King, S. (2013). Noise robustness in hmm-tts speaker adaptation. In *Eighth ISCA Workshop on Speech Synthesis.*

Yang, B., Zou, Y., Liu, F., and Zhang, C. (2021). Non-autoregressive coarse-to-fine video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3119–3127.

Yeung, J. M. C. (1974). *Towards spoken English: a computer based synthesizer for a reading machine for the blind.* PhD thesis, University of British Columbia.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2001). Mixed excitation for hmm-based speech synthesis. In *Seventh European conference on speech communication and technology*.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2005). Incorporating a mixed excitation model and postfilter into hmm-based text-to-speech synthesis. *Systems and Computers in Japan*, 36(12):43–50.

Young, N. (1957). Some factors affecting intelligibility in single sideband communications. *IRE Transactions on Communications Systems*, 5(1):96–98.

Yu, A. C. (2022). Perceptual cue weighting is influenced by the listener's gender and subjective evaluations of the speaker: the case of english stop voicing. *Frontiers in Psychology*, 13:840291.

Yu, C., Fu, C., Chen, R., and Tapus, A. (2022). First attempt of gender-free speech style transfer for genderless robot. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1110–1113. IEEE.

Yu, J., Zhang, M., Tao, J., and Wang, X. (2007). A novel hmm-based tts system using both continuous hmms and discrete hmms. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–709. IEEE.

Yu, Y., Zhu, F., Li, X., Liu, Y., Zou, J., Yang, Y., Yang, G., Fan, Z., and Wu, X. (2013). Overview of shrc-ginkgo speech synthesis system for blizzard challenge 2013. In *Blizzard Challenge Workshop*, volume 2013.

Zahorian, S. A. (1979). Principal-components analysis for low redundancy encoding of speech spectra. *The Journal of the Acoustical Society of America*, 65(4):1069–1069.

Zen, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pages 7962–7966. IEEE.

Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

Zhao, Z. Z. Y. G. X. Y. X. (2020). Synspeechddb: a new synthetic speech detection database.

Zsiga, E. C. (2013). *The sounds of language: An introduction to phonetics and phonology*, volume 7. John Wiley & Sons.

Łańcucki, A. (2021). Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592.