Tampere University

Elmeri Grönlund

INFANT-DIRECTED SPEECH SYNTHESIS WITHOUT IN-DOMAIN TRAINING DATA WITH A ZERO-SHOT TEXT-TO-SPEECH MODEL

Master's thesis
Faculty of Information Technology and Communication Sciences
Examiners: Okko Räsänen, Daniil Kocharov
May 2026

ABSTRACT

Elmeri Grönlund: Infant-Directed Speech Synthesis without In-Domain Training Data with a Zero-Shot Text-To-Speech Model
Master's thesis
Tampere University
Master's Programme in Information Technology
May 2026

Infant-directed speech (IDS) is a distinctive manner of speaking used when addressing infants and young children, and it has been suggested to support infants' early language development. Recent work has proposed that IDS could be used as input to computational models of early language acquisition, enabling controlled simulation studies of early language development. However, recordings of IDS are typically noisy or otherwise of non-ideal quality due to uncontrolled recording conditions in natural environments. There are also ethical considerations related to the use of privacy-sensitive real-world data from infants' everyday interactions. Synthetic IDS has been proposed as an alternative to real recordings; however, no text-to-speech (TTS) system currently produces high-quality, realistic IDS. This is primarily because neural TTS models require high-quality training data, while publicly available datasets for IDS are often of insufficient quality. To address these limitations, the present study investigates whether a reference-based zero-shot neural TTS system can produce intelligible, style-consistent IDS when using natural IDS as reference speech.

The original IDS recordings were first enhanced using a neural speech enhancer to improve their audio quality. The enhanced reference speech was then paired with IDS transcripts in order to synthesize the text in the style of the reference speech. The transcripts were generated by a neural language model designed to capture the simplified linguistic structures typical of IDS.

The TTS and enhancement systems were evaluated for their ability to preserve key prosodic characteristics of IDS, including mean and variability of fundamental frequency, as well as speaking rate. Results indicated that the system adequately transferred F0-related statistics and produced intelligible speech. Neural enhancement had a slight positive effect on prosodic transfer from reference to synthetic speech and the intelligibility of the synthesized speech. However, certain features, such as syllable duration, were not accurately preserved, revealing limitations in the investigated TTS model's ability to transfer prosody from the source signal.

Overall, the findings demonstrate that reference-encoder-based neural TTS can synthesize IDS from naturalistic recordings with reasonable fidelity, without necessitating the use of models trained directly on source IDS data. The study highlights the potential and current constraints of the approach, providing a promising ground for efficiently creating high-quality synthetic IDS.

Keywords: infant-directed speech, speech, speech synthesis, text-to-speech, acoustic features, prosody, intelligibility, neural network, enhancement, denoising

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

TIIVISTELMÄ

Elmeri Grönlund: Vauvoille suunnatun puheen synteesi käyttäen tyylilliseen adaptoitumiseen kykenevää neuraalista syntesaattoria
Diplomityö
Tampereen yliopisto
Tietotekniikan DI-ohjelma
Toukokuu 2026

Vauvoille suunnattu puhe (IDS, Infant-Directed Speech) on erityinen puhetapa, jota käytetään vauvoille ja pienille lapsille puhuttaessa, ja sen on esitetty tukevan lasten varhaista kielenkehitystä. Viimeaikainen tutkimus on ehdottanut, että IDS:ää voitaisiin käyttää syötteenä varhaisen kielenomaksumisen laskennallisissa malleissa, mahdollistaen kontrolloituja simulaatioita varhaisen kielenkehityksen tutkimiseksi. IDS-äänitteet ovat kuitenkin tyypillisesti kohinaisia tai muutoin heikkolaatuisia johtuen puhetyylin luonnollisten tallennusympäristöjen kontrolloimattomista olosuhteista. Lisäksi lapsiperheiden arkisista vuorovaikutustilanteista nauhoitettu aineisto on yksityisyyden suojan kannalta herkkää, ja sen käyttöön liittyy useita eettisiä kysymyksiä. Synteettinen IDS on siten esitetty vaihtoehdoksi todellisille äänitteille; kuitenkaan yksikään tekstistä puheeksi -järjestelmä (TTS, text-to-speech) ei tällä hetkellä ole osoittautunut kykeneväksi tuottamaan korkealaatuista ja luonnollista IDS:ää, pääasiassa siksi, että neuroverkkoihin perustuvat TTS-mallit edellyttävät korkealaatuista opetusdataa, ja julkisesti saatavilla olevat aineistot ovat usein laadultaan heikkoja ja siten riittämättömiä.

Tässä tutkimuksessa selvitetään, voisiko referenssi-enkooderipohjainen zero-shot TTS-järjestelmä tuottaa ymmärrettävää ja tyyliltään johdonmukaista IDS:ää, kun luonnollisia IDS-äänitteitä käytetään referenssipuheena, josta IDS-tyylin akustisia piirteitä siirretään synteettiseen puheeseen. Alkuperäisiä IDS-äänitteitä ehostettiin neuroverkkoon perustuvalla puheen ehostusjärjestelmällä niiden laadun parantamiseksi. TTS-mallin syötteenä käytettiin ehostetun referenssipuheen lisäksi IDS-tyylille ominaista kielisisältöä sisältäviä tekstejä. Tekstit tuotettiin kielimallilla, joka erikoistuu IDS:lle tyypillisten yksinkertaistettujen kielirakenteiden tuottamiseen.

TTS- ja puheen ehostusjärjestelmiä arvioitiin sen perusteella, kuinka hyvin ne säilyttävät IDS:n keskeiset prosodiset piirteet. Piirteisiin lukeutuvat esimerkiksi fundamentaalisen taajuuden (F0) taso ja vaihtelu sekä puhenopeus. Tulokset osoittivat, että IDS-referensseihin ehdollistettuna järjestelmä kykeni siirtämään F0:aan liittyviä tilastollisia ominaisuuksia ja tuotti ymmärrettävää puhetta. Puheen ehostuksella oli lievä positiivinen vaikutus prosodian siirtymiseen referenssipuheesta synteettiseen puheeseen sekä synteettisen puheen ymmärrettävyyteen. Joitakin piirteitä, kuten tavunkestoja, ei kuitenkaan siirtynyt johdonmukaisesti referenssistä synteesiin, mikä paljasti nykyisen TTS-mallin rajoituksia.

Tulokset osoittavat, että referenssi-enkoodereihin perustuva zero-shot TTS-järjestelmä pystyy tuottamaan IDS:ää luonnollisista äänitteistä kohtuullisella tarkkuudella ilman, että mallia tarvitsee kouluttaa suoraan IDS-aineistolla. Tutkimus korostaa lähestymistavan mahdollisuuksia ja nykyisiä rajoituksia sekä tarjoaa lupaavan perustan korkealaatuisen synteettisen IDS:n tehokkaalle tuottamiselle.

Avainsanat: vauvoille suunnattu puhe, puhe, puhesynteesi, text-to-speech, akustiset piirteet, prosodia, ymmärrettävyys, neuroverkko, ehostus, kohinan poisto

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

Use of artificial intelligence in the thesis

Artificial intelligence applications have been used in my thesis:

☐ No

☒ Yes

According to my declaration, I have used the following artificial intelligence applications during the thesis process:

Names and versions of AI applications: GPT-4.1, DALL-e v2

Purpose of use: Language editing. Correction of grammar and writing style. Assistance in image creation.

Sections where AI was used:

AI was used for language revision throughout the entire thesis.

DALL-e V2 was used to generate the surrounding anatomical context for Figures 2.1 and 2.2.

I acknowledge that I am fully responsible for the entire content of my thesis, including the parts where artificial intelligence has been utilized, and I accept responsibility for any potential violations of ethical guidelines.

CONTENTS

1. Introduction . . . 1
2. Background . . . 3
2.1 Speech . . . 3
2.1.1 Speech Production, Language Encoding, and Speech Perception . . . 4
2.1.2 Acoustic Properties of Speech . . . 6
2.1.3 Intelligibility . . . 7
2.1.4 Analysis of Speech Signals . . . 8
2.2 Styles of Speech and Infant-Directed Speech . . . 11
2.2.1 Read and Spontaneous Speech . . . 12
2.2.2 Infant-Directed Speech . . . 13
2.3 Non-Neural Approaches to Text-to-Speech . . . 15
2.3.1 Classical source–filter synthesis approaches . . . 15
2.3.2 Concatenative and Statistical parametric synthesis . . . 16
2.4 Neural Approaches to Text-to-Speech Synthesis . . . 18
2.4.1 Sequence-to-Sequence Architectures . . . 19
2.4.2 Convolutional Neural Networks in Sequence-to-Sequence TTS . . . 20
2.4.3 Transformers and Transformer-Based Architectures . . . 21
2.4.4 Neural Vocoders . . . 23
2.5 Speaker and Style Control in Neural TTS . . . 24
2.5.1 Speaker Identity and Speech Style . . . 25
2.5.2 Multi-Speaker and Style Embeddings . . . 25
2.5.3 Zero-Shot Speaker Adaptation . . . 26
2.5.4 Text-Driven Style Control . . . 26
2.6 Preprocessing Speech Data . . . 27
2.6.1 Noise Reduction and Audio Enhancement . . . 27
2.6.2 Data Quality and Quantity Requirements for Neural TTS Systems . . . 29
2.7 Technological Applications of IDS Modeling . . . 31
3. Methods . . . 33
3.1 Data and the Text-to-Speech pipeline . . . 33
3.1.1 ManyBabies 1 Corpus . . . 34
3.1.2 TIMIT . . . 35
3.1.3 Text Prompts for Synthesis . . . 35
3.1.4 Preprocessing with Resemble Enhance . . . 36
3.1.5 XTTSv2 for Zero-Shot Style Transfer . . . 37

3.2 Analysis of Acoustic Features and Intelligibility . . . 39
3.2.1 Segmentation into Utterances . . . 39
3.2.2 Acoustic and Intelligibility Measures. . . . 40
3.3 Experimental Design . . . 41
3.3.1 Effect of Style on Prosody Transfer from Reference to Synthesis . . . 42
3.3.2 Effect of Preprocessing on Prosody Transfer from Reference to Synthesis . . . 42
3.3.3 Effect of Text Prompt Style on Prosody Transfer from Reference to Synthesis . . . 43
3.3.4 Effect of Reference Concatenation on Prosody Transfer from Reference to Synthesis . . . 43
3.3.5 Intelligibility of the Synthesized Speech . . . 43
4. Results . . . 45
4.1 Transfer of Prosody from Reference to Synthesis . . . 45
4.1.1 Effect of Reference Style . . . 45
4.1.2 Effect of Reference Preprocessing . . . 47
4.1.3 Effect of Text Prompt Style . . . 51
4.1.4 Effect of Reference Concatenation . . . 52
4.2 Intelligibility of the synthesized speech . . . 53
5. Discussion . . . 57
5.1 Transfer of IDS-related acoustic characteristics . . . 57
5.1.1 Effect of reference audio enhancement on the transfer of prosody . . . 58
5.1.2 Effect of text prompt style and reference audio concatenation on prosody transfer . . . 59
5.2 Intelligibility of the Synthesis . . . 59
5.3 Study Limitations . . . 60
5.4 Ethical considerations . . . 61
5.5 Future work . . . 62
6. Conclusion . . . 64
References . . . 65

LIST OF FIGURES

2.1 Simplified physiological overview of human speech production. 5
2.2 Simplified physiological overview of human auditory speech perception. . . 6
2.3 Visualized speech feature representations from waveform to mel-frequency cepstral coefficients (MFCC). 11
2.4 Simplified vocal tract tube model 17
2.5 Pepper robot interacting with a child in a mall (2016). 31

3.1 Example of text prompts generated by GILES and ChatGPT. 36
3.2 xTTS-v2 architecture overview. 38
3.3 The TTS pipeline and analysis procedures of the study 39
3.4 The evaluation framework of the study 41

4.1 Scatter plots with linear fit comparing original reference speech and non-enhanced synthesis (IDS text prompts). 46
4.2 Crosswise comparison of synthesized speech from IDS and ADS references prompted with IDS texts (GILES). 47
4.3 KDEs of prosodic and spectral features across synthesis conditions. . . . 48
4.4 Scatter plots comparing original and enhanced IDS and ADS references. . 50
4.5 Scatter plots comparing original reference speech and synthesis from enhanced references. 51
4.6 Scatter plot with linear fit comparing IDS and ADS reference speech with synthesized outputs (ADS text prompts). 52
4.7 Scatter plots comparing IDS synthesis from non-concatenated and concatenated inputs. 53
4.8 Original vs enhanced ManyBabies 1 (MB 1) word error rate (WER) distributions. 54
4.9 WER distributions of synthesized speech across reference and text conditions. 55
4.10 Mean WER across speech corpora and synthesis conditions in this study. 55

LIST OF TABLES

3.1 IDS and ADS duration statistics per speaker (MB 1 subset) 35
3.2 Duration statistics of the subset of TIMIT dataset. 35

4.1 Jensen–Shannon distances between kernel-density estimations across pre-processing conditions. 49
4.2 Aggregated statistically significant correlations between reference and synthesized speech in IDS and ADS text styles (GILES vs. QTEXT). 52

LIST OF SYMBOLS AND ABBREVIATIONS

ADS	adult-directed speech
ASR	automatic speech recognition
CDS	child-directed speech
concatenation	The process of joining two or more speech segments end-to-end to form a continuous signal
DCT	Discrete Cosine Transform used for feature decorrelation
F0	Fundamental frequency of speech, corresponding to perceived pitch
formant	Resonance frequency of the vocal tract shaping speech sounds
frame	Short segment of signal used for spectral analysis
HMM	hidden Markov model
IDS	infant-directed speech
intelligibility	A measure of how accurately spoken content can be perceived and understood by a listener
KDE	kernel density estimation
LP	linear prediction
MB 1	ManyBabies 1
Mel scale	Perceptual frequency scale based on human hearing
MFCC	mel-frequency cepstral coefficients
paralinguistic	Non-verbal aspects of speech that convey information beyond linguistic content.
prosody	Rhythm, stress, and intonation patterns of speech
Resemble Enhance	A neural speech enhancement system that improves speech quality by reducing noise and artifacts while preserving speaker identity and linguistic content
source–filter	A speech production model where speech is viewed as the output of a source excitation (e.g., glottal pulses or noise) passed through a vocal tract filter that shapes the spectral envelope
spectral envelope	Smooth curve outlining spectral peaks of a sound

spectral tilt	Slope of speech spectrum indicating energy distribution across frequencies
spectrogram	Time-frequency representation of a signal
STFT	Short-Time Fourier Transform
TAU	Tampere University
TIMIT	Phonetically balanced speech corpus for ASR research
TTS	text-to-speech
TUNI	Tampere Universities
URL	Uniform Resource Locator
utterance	A continuous segment of acoustic speech signal bounded by periods of silence or low energy.
vocoder	An analysis–synthesis system that can decompose an acoustic signal into a parametric representation and reconstruct a time-domain waveform from the parameters
WER	word error rate
Whisper	Neural automatic speech recognition system by OpenAI
window function	A function applied to a signal to reduce spectral leakage
xTTS-v2	Cross-lingual zero-shot text-to-speech system

1. INTRODUCTION

Infant-directed speech (IDS) is a specific manner of speech used when addressing infants and young children, and it systematically differs from adult-directed speech (ADS) in several acoustic and prosodic dimensions (Snow 1972; Soderstrom 2007). Infants have been found to consistently prefer IDS over ADS, attending to it longer and showing stronger responses to it (Fernald 1985; Frank et al. 2020). These findings suggest that IDS is particularly effective in engaging infants and may support early language acquisition.

Interest in computational modeling of early language acquisition has grown in recent years (see Benders and Blom 2023; Brent 1999; Kuhl 2004; Räsänen 2012 for reviews). Consequently, researchers have increasingly focused on the role of IDS in computational models of language development (Lavechin et al. 2024; Räsänen, Cruz Blandón et al. 2024). However, using IDS data collected directly in natural environments for modeling purposes is not straightforward.

Most existing large datasets of IDS consist of audio recordings collected in natural home environments using child-worn recording devices (Bergelson et al. 2023; Demuth et al. 2006; Sullivan et al. 2021). Such data present challenges not only in terms of the ethical considerations involved in processing sensitive recordings (Cychosz, Romeo et al. 2020), but also in terms of technical robustness. The audio quality of day-long recordings is often degraded by uncontrolled recording conditions. Home environments typically introduce non-stationary background noise, such as overlapping speech from the child and other household sounds (e.g., television, appliances, and movement within the environment). The dynamic nature of these settings also introduces technical issues, including changes in speaker–microphone distance and acoustic conditions shaped by differences in room size, layout, and reverberation as the child moves between rooms. Because wearable microphones must not constrain the child, compromises in placement, sensitivity, and form factor are required. As a result, the audio is typically noisier, less consistent, and more prone to distortion than recordings made in controlled studio environments. In contrast, IDS recordings collected in laboratory or semi-controlled settings are typically of higher quality, but the amount of speech per speaker is relatively limited. Such controlled settings, while capable of eliciting valid IDS (Frank et al. 2020), can also affect the acoustic and linguistic properties of natural IDS.

One potential approach to addressing the limitations of existing IDS datasets is the use of synthetic IDS. However, while the development of deep learning has led to rapid advances in text-to-speech (TTS) systems (Casanova, Davis et al. 2024; Lameris et al. 2023; Oord, Dieleman et al. 2016; Ren et al. 2019; Wang, Skerry-Ryan et al. 2017), and modern systems can now generate high-quality ADS, no TTS model currently exists that is specifically designed to produce IDS. One reason for this is the previously described limitations of IDS datasets. Training a neural TTS model capable of reliably generating IDS would require high-quality IDS audio, which is currently not available in sufficient quantities.

In this thesis, the possibility of using a TTS model for IDS-style speech generation is investigated. The focus is on a TTS model capable of synthesizing speech that matches the voice characteristics of a previously unseen speaker from a short reference audio sample. Although the TTS model is able to transfer speaker voice characteristics, the study focuses solely on transferring the speaking style of the reference to the synthetic speech.

Prior work has reported difficulties in training TTS models with noisy data (Zhang, Ren, Deng et al. 2023). In addition, preliminary experiments in this study indicate that noisy reference inputs can cause issues during inference, specifically due to the model's tendency to reproduce signal-level artifacts present in lower-quality inputs. As most naturally recorded IDS is inherently noisy, the TTS system is combined with neural speech enhancement to mitigate these effects. This setup allows investigating whether audio enhancement improves IDS prosody transfer from reference to synthesis as well as its effect on the intelligibility of the generated speech. The resulting pipeline aims to transform non-ideal-quality IDS recordings into higher-quality synthesized speech while preserving the style-related acoustic characteristics of the reference.

This study does not examine whether the synthesized speech preserves the language-learning benefits of IDS from a psycholinguistic perspective, as addressing this question would require a separate, dedicated investigation. Instead, the focus is on generating synthetic speech that preserves the acoustic properties and qualitative features of natural IDS (Cox et al. 2023; Martin et al. 2016; Miyazawa et al. 2017; Soderstrom 2007), without training generative models directly on real-world IDS data.

2. BACKGROUND

This chapter provides the background necessary to understand modern text-to-speech (TTS) systems and their application to infant-directed speech (IDS). In the beginning an overview of human speech is given, including physical production and perception, as well as key acoustic properties and common analysis techniques used to characterize speech signals (Section 2.1). Based on this foundation, the concept of speech styles is then introduced from an acoustic and signal analysis perspective (Section 2.2).

The chapter then reviews classical (Section 2.3) and modern neural approaches to speech synthesis. In particular, neural approaches are further divided into general text-to-speech synthesis (Section 2.4) and methods that incorporate explicit control over speaker identity and speaking style (Section 2.5).

Finally, the chapter discusses challenges faced by neural TTS systems, particularly those related to data quality, preprocessing, and robustness to variability in recording conditions and speaker characteristics (Section 2.6). This is followed by possible applications of text-to-IDS (Section 2.7).

2.1 Speech

Human speech is a complex multimodal phenomenon that can be examined from linguistic, physiological, cognitive, and acoustic perspectives. At a physical level, speech consists of variations in air pressure that propagate as sound waves, within which linguistic information (e.g., words) and paralinguistic information (e.g., prosody) are encoded (Bäckström et al. 2022, Ch. 2). This information can be decoded not only as a communicative message, but also in terms of intent, identity, emotion, and interpersonal stance (Bäckström et al. 2022, Sec. 8.6). These communicative goals shape both speech production and perception. The resulting acoustic signal integrates these layers of information, which listeners interpret using perceptual mechanisms and prior knowledge. Consequently, the development of artificial speech systems and the analysis of their outputs must consider how communicative intentions are mapped onto acoustic signals, how these signals are interpreted, and, in some cases, how speech is produced by the human vocal system.

This section provides an overview of speech as a concept and outlines key consider-

ations involved in modeling or artificially reproducing human speech. First, Subsection 2.1.1 presents human speech production and also considers speech from the listener's perspective in terms of both physical and perceptual processing. Subsections 2.1.2 and 2.1.3 then provide an overview of the central acoustic properties relevant to this study. Finally, Subsection 2.1.4 reviews common methodologies for speech signal analysis and feature extraction.

2.1.1 Speech Production, Language Encoding, and Speech Perception

Speech is produced as air is expelled from the lungs and shaped by the human vocal system, as shown in Figure 2.1. As airflow passes through the larynx, vibration of the vocal folds may occur, producing a periodic acoustic source corresponding to voiced speech sounds (Lieberman and Blumstein 1988, pp. 3–13). In the absence of vocal fold vibration, an aperiodic or noise-like source is generated instead. Above the larynx, this source is shaped by the vocal tract, which acts as a dynamic acoustic filter defined by the pharyngeal, oral, and nasal cavities (Lieberman and Blumstein 1988, pp. 3–13). The resonant properties of this system are continuously modified by articulatory movements of the tongue, lips, jaw, and velum (Ladefoged and Johnson 2011, pp. 2–17). The velum in particular controls whether the nasal cavity is acoustically coupled to the vocal tract. The resulting filtered acoustic signal is ultimately radiated through the lips and/or nasal cavity, forming the continuous speech waveform.

Within the articulatory system, speech sounds vary according to how and where airflow is shaped by the vocal tract. Vowels are produced with a relatively open vocal tract and differ primarily in tongue height, tongue advancement, and lip rounding, which together determine their acoustic quality (Ladefoged and Johnson 2011, pp. 8–21). Consonants involve greater constriction of airflow and are distinguished by place of articulation (e.g., bilabial, alveolar, velar), manner of articulation (e.g., stops, fricatives, nasals), and whether vocal fold vibration occurs, yielding voiced and voiceless contrasts such as */b/* versus */p/* (Ladefoged and Johnson 2011, pp. 33–45, 56–68). The articulatory parameters vary along continuous dimensions rather than as fixed categories, and their movements overlap in time. This overlap produces continuous speech, where phenomena such as coarticulation occur (Ladefoged and Johnson 2011, pp. 70–72), in which the production of a sound is influenced by surrounding sounds. For example, the articulation of a vowel may shift slightly depending on adjacent consonants.

Language provides a systematic framework that maps communicative meaning in the continuous waveform of sound onto discrete symbolic representations (Fromkin et al. 2018, pp. 394–415). Languages select, categorize, and organize subsets of the available articulatory space into conventional sound patterns and combinations (Fromkin et al.

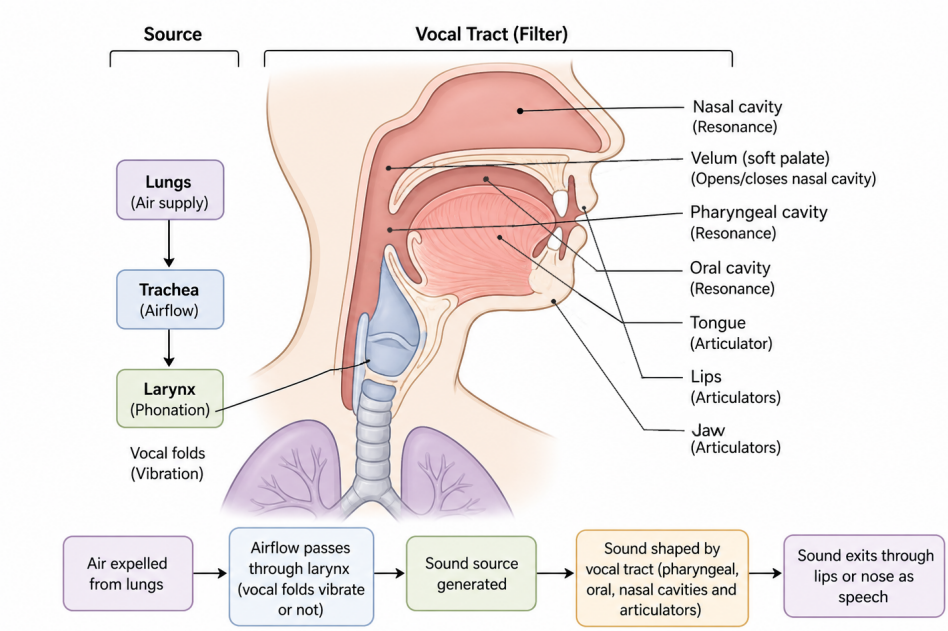

Figure 2.1. *Simplified physiological overview of human speech production. The source component includes the lungs (air supply), trachea (airflow), and larynx, where phonation occurs via possible vocal fold vibration. The vocal tract component includes the pharyngeal cavity, oral cavity, and nasal cavity, which act as resonant spaces shaping the sound. The velum (soft palate) controls coupling between the oral and nasal cavities. Articulation is performed by the tongue, lips, and jaw, which modify the shape of the vocal tract to produce different speech sounds. Sound radiation occurs through the lips and/or nasal cavity as the acoustic signal exits the vocal system. AI was used to assist in the creation of this image by generating the surrounding anatomical context (DALL·E v2).*

2018, pp. 1–8). These systems differ across linguistic communities, reflecting historical, social, and cultural influences. Through learning, speakers acquire shared conventions that map articulatory patterns onto linguistic units of varying complexity, for example from individual phonemes to sequences of phonemes forming words, and from words to sequences forming sentences.

Speech must also be perceived acoustically to fulfill its communicative function. When speech reaches the ear, sound waves enter the outer ear and induce vibrations of the tympanic membrane (eardrum), as illustrated in Figure 2.2. These vibrations are transmitted via the ossicles in the middle ear to the cochlea, where they are converted into electrical signals by hair cells along the basilar membrane(see Oxenham 2018 for a review). Different frequency components activate distinct regions of the cochlea, enabling a tonotopic representation of the speech signal in which frequency is systematically mapped onto spatial locations along the cochlea (Oxenham 2018).

The auditory nerve carries these signals to the brain, where they are processed in the auditory cortex. The auditory system integrates temporal, spectral, and amplitude cues

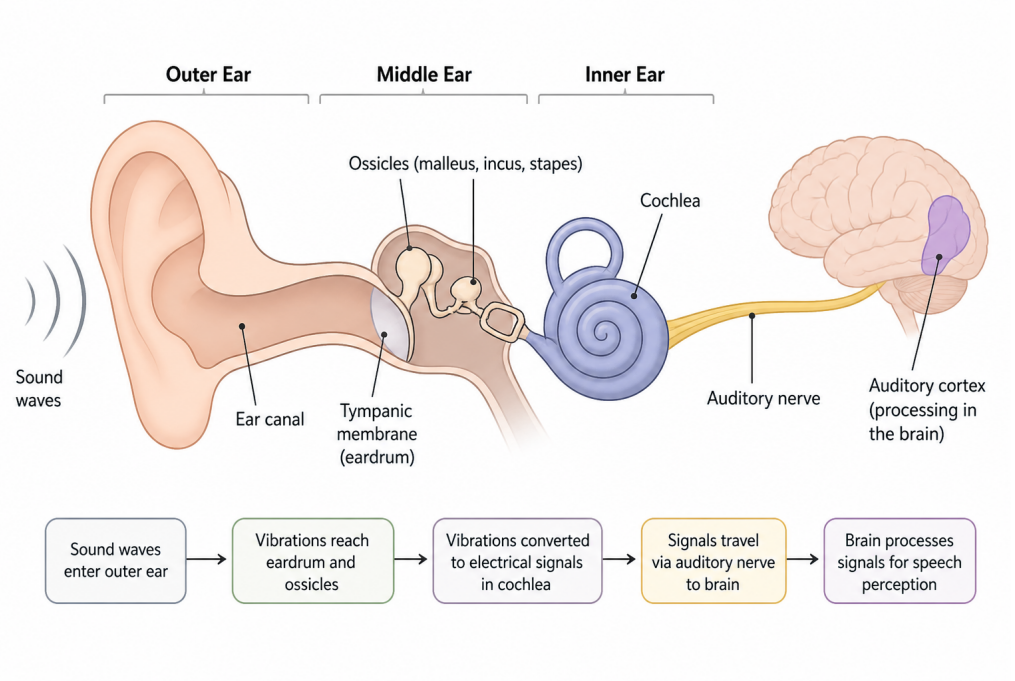

Figure 2.2. *Simplified physiological overview of human auditory perception showing the main stages involved in processing acoustic signals, including speech. The outer ear receives incoming sound waves and directs them through the ear canal toward the tympanic membrane (eardrum). In the middle ear, vibrations are mechanically transmitted and amplified by the ossicles, consisting of the malleus (hammer), incus (anvil), and stapes (stirrup), before reaching the inner ear. The cochlea converts these mechanical vibrations into electrical neural signals, which are transmitted via the auditory nerve to the brain, where auditory cortical regions process the incoming information for speech perception. AI was used to assist in the creation of this image by generating the surrounding anatomical context (DALL·E v2)*

to recover phonetic structure and suprasegmental features such as intonation, rhythm, and stress (Oxenham 2018). This interpretation relies on both sensory input and prior knowledge, including language-specific expectations, lexical memory, and contextual information. Linguistic and paralinguistic information are not decoded independently; for instance, variations in pitch may signal sentence modality, emphasis, emotional state, or speaker intent depending on context (Themistocleous 2025).

2.1.2 Acoustic Properties of Speech

Speech is perceived as a unified sound, but it can be analyzed in terms of distinct yet interacting acoustic properties. One fundamental property is periodicity, which arises from the regular vibration of the vocal folds discussed in Section 2.1.1. This vibration generates a harmonic spectrum with peaks at the F0 and its integer multiples. The rate of vocal fold vibration determines F0, which is perceived as pitch (Bäckström et al. 2022, Ch. 2; Naylor 2014). Variations in F0 over time form intonation patterns, which convey

both linguistic functions related to information structure (e.g., emphasis and sentence structure) and paralinguistic information such as affect and emotional arousal (Bäckström et al. 2022, Ch. 2; Bänziger and Scherer 2005; Cox et al. 2023).

In addition to periodicity, spectral structure provides important information about a speaker's voice and phonetic identity. The vocal tract acts as a resonant filter, amplifying certain frequency bands to produce formant peaks in the spectrum that are crucial for distinguishing vowels, as each vowel is characterized by a specific pattern of formant frequencies. The resulting formant configuration forms the spectral envelope, which describes the coarse spectral shape of speech.

A related spectral feature is spectral tilt, which describes how energy is distributed between low and high frequencies. For a frame with magnitude spectrum $X(f)$, spectral tilt T can be estimated as the slope of a linear fit:

$$\log_{10} |X(f)| \approx T \cdot f + b, \tag{2.1}$$

where a steeper negative slope indicates relatively more low-frequency energy, and a flatter slope indicates relatively more high-frequency energy. Different values of spectral tilt are associated with perceptual voice quality differences. Spectra dominated by low-frequency energy (steeper negative tilt) are typically perceived as softer, breathier, or more relaxed, whereas flatter tilt emphasizing higher frequencies is associated with tenser, sharper, or more pressed voice quality (Garellek et al. 2016; Hillenbrand et al. 1994).

Another key property is amplitude, which reflects sound intensity. Amplitude depends on airflow through the vocal folds and their configuration and contributes to perceived loudness (Bäckström et al. 2022, Ch. 2). Speakers can use amplitude variations to emphasize words or convey prominence. Amplitude also helps delineate speech from silence and contributes to the temporal structure of phonetic segments.

Timing governs the duration of sounds, the spacing between them, and the rhythm of utterances. An utterance is typically defined as a continuous stretch of acoustic activity separated by silence or low-amplitude regions, which does not necessarily correspond to grammatical sentence boundaries (Bäckström et al. 2022, Ch. 2). Speech rate, often measured in syllables or words per second, provides a quantitative description of temporal patterns, and utterance-level timing allows for consistent comparison across recordings.

2.1.3 Intelligibility

Speech conveys communicative intent, but its meaning is only realized when it is interpreted by a listener. Intelligibility refers to how accurately a listener can perceive and understand the intended linguistic content, and it reflects the interaction between acous-

tic cues and perceptual and cognitive processes (Bradlow et al. 1996; Ishikawa et al. 2021). It is influenced by properties such as speech rate, articulation clarity, and prosody, as well as by external conditions related to overall audio quality, including background noise, reverberation, overlapping speech, and transmission artifacts.

Importantly, intelligibility is shaped not only by the acoustic signal itself but also by the listener. Human listeners are generally adaptive and use linguistic knowledge, expectations, and familiarity with a speaker or language to interpret degraded signals (Chen 2009). Speech can also be processed by artificial systems such as TTS and automatic speech recognition (ASR), which can be viewed as machine counterparts to human listeners. Machine perception differs from human perception: for human listeners, intelligibility and perceived naturalness are central, whereas ASR systems are primarily sensitive to acoustic fidelity. As a result, distortions that are tolerable for humans may significantly degrade recognition performance. In particular, background noise, altered prosody, or atypical speaking styles can reduce ASR accuracy, thereby affecting downstream applications such as voice-controlled assistants (Chen 2009).

Consequently, intelligibility can be assessed from both human and machine perspectives. Human evaluation is typically carried out through subjective listening tests, where participants directly judge the intelligibility of the speech they hear. While these tests provide the most direct measure of perceived performance, they are costly to conduct and may exhibit variability due to differences in listener experience, context, and task design (Bäckström et al. 2022, Ch. 6). To improve comparability and structure, standardized protocols such as mean opinion score (MOS) are commonly used to quantify subjective judgments in a reproducible manner (Bäckström et al. 2022, Ch. 6).

In contrast, objective evaluation methods approximate human judgments using computational models. Perceptual quality or intelligibility can be estimated using these models. For example, one could derive a transcription of the speech using ASR and compute the word error rate (WER) by comparing it with a ground-truth transcription. Although objective metrics are ultimately approximations of human perception, they are widely used due to their efficiency, repeatability, and suitability for the development and benchmarking of speech processing systems (Bäckström et al. 2022, Ch. 6).

2.1.4 Analysis of Speech Signals

To study speech systematically the acoustic waveform can be analyzed at multiple levels. Certain features, such as amplitude, timing and duration, can be measured directly from the waveform in the time domain. These measurements provide valuable information about the temporal structure of speech including syllable duration pauses and rhythm (Bäckström et al. 2022, Ch. 3). However, many important properties of speech, such as pitch harmonic structure formants and the spectral envelope, require examining the signal

in the frequency domain (Bäckström et al. 2022, Ch. 3). Transforming the waveform into frequency based representations allows detailed analysis of how energy is distributed across frequencies and how it evolves over time.

Speech signals vary continuously and cannot be described by a single global measure. To capture temporal dynamics while also preparing for spectral analysis the waveform is typically divided into short overlapping segments called frames. Within each frame, the signal can be approximated as being quasi-stationary, meaning its acoustic properties such as amplitude and frequency content do not change significantly over the short segment (Bäckström et al. 2022, Ch. 3). This approximation enables spectral and temporal features to be extracted from each frame to characterize its contents.

Dividing the continuous signal into frames introduces discontinuities at the edges of each segment, which can cause spectral leakage where energy from one frequency spreads into neighboring frequencies (Bäckström et al. 2022, Ch. 3). To mitigate this, each frame is multiplied by a smooth window function, a process known as windowing (Harris 1978; Bäckström et al. 2022, Ch. 3). If $x[n]$ is the n-th sample of a frame of length N, the windowed signal $x_w[n]$ is

$$x_w[n] = x[n] \cdot w[n], \quad n = 0, 1, \ldots, N-1, \tag{2.2}$$

where $w[n]$ is the window function. Common examples include the Hamming window (Bäckström et al. 2022, Ch. 3):

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \tag{2.3}$$

and the Hann window

$$w[n] = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right). \tag{2.4}$$

Windowing tapers the edges of each frame, reducing discontinuities and spectral leakage (Bäckström et al. 2022, Ch. 3). Typical window lengths for speech analysis are 20 to 40 milliseconds. Shorter frames improve the stationarity assumption but reduce frequency resolution, while longer frames improve frequency resolution but may violate stationarity. Therefore, the choice of frame length represents a trade-off between time and frequency resolution and often depends on the analysis goal. In practice frames often overlap by 50 percent to improve temporal continuity and reduce information loss at frame boundaries.

Once a signal has been framed and windowed, its frequency content can be analyzed using the Short-Time Fourier Transform (STFT). The STFT describes how the spectral content of a signal evolves over time by applying a Fourier transform to localized segments of the signal (Bäckström et al. 2022, Ch. 3).

For a continuous-time signal $x(t)$, the Fourier transform is defined as (Bäckström et al. 2022, Ch. 3)

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft}\, dt, \tag{2.5}$$

and the short-time Fourier transform is obtained by applying this transform to a windowed portion of the signal (Bäckström et al. 2022, Ch. 3)

$$X(\tau, f) = \int_{-\infty}^{\infty} x(t)\, w(t-\tau)\, e^{-j2\pi ft}\, dt, \tag{2.6}$$

where $w(t-\tau)$ is a window function centered at time τ.

Typically, in digital signal processing, speech signals are represented in discrete time and processed in short, overlapping frames. For a discrete-time segment consisting of N samples $x[n]$, the frequency-domain representation of each frame is commonly obtained using the Discrete Fourier Transform (DFT) (Bäckström et al. 2022, Ch. 3):

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}, \quad k = 0, 1, \ldots, N-1. \tag{2.7}$$

Thus, in practice the STFT of a discrete signal can be implemented by applying the DFT to successive windowed frames.

The Fourier transform decomposes a signal into sinusoidal components, producing complex values that encode both magnitude $|X[k]|$, representing the spectral energy distribution, and phase $\angle X[k]$, representing the relative timing of the frequency components (Bäckström et al. 2022, Ch. 3). In many speech analysis tasks, only the magnitude spectrum is retained. However, phase information is essential when accurately reconstructing a waveform from a spectrogram.

Spectrograms are computed from the STFT of overlapping frames (Allen and Rabiner 1977; Bäckström et al. 2022, Ch. 3). Perceptually motivated variants such as the logarithmic Mel scale spectrum (Bäckström et al. 2022, Ch. 3) apply a bank of overlapping triangular filters to the magnitude spectrum with center frequencies spaced according to the Mel scale and converted to logarithmic. This scale reflects the nonlinear frequency resolution of the human auditory system where low-frequency components are represented with higher precision while higher frequencies are progressively compressed.

Speech can also be represented in the cepstrum. The cepstrum is obtained by taking the inverse fourier transform of the logarithm of the magnitude spectrum (Bäckström et al. 2022, Ch. 3). In practice, the short-time magnitude spectrum is first filtered using a Mel scale filter bank, and then a DCT is applied to the log-magnitude outputs to produce the widely used MFCCs (Davis and Mermelstein 1980; Bäckström et al. 2022, Ch. 3) The

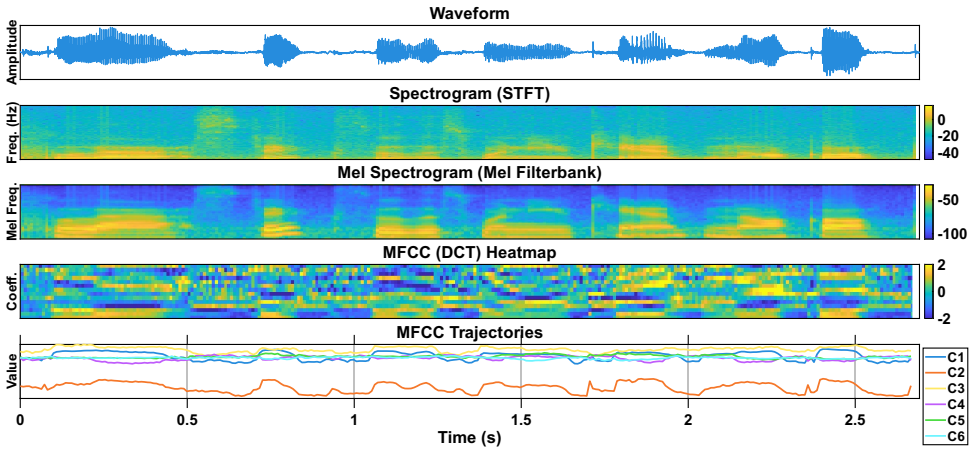

Figure 2.3. *The figure illustrates visualized representations of a speech signal. The displayed mel-frequency cepstral coefficients (MFCC) heatmap is shown after mean–variance normalization across time for visualization purposes. Individual MFCC trajectories correspond to cepstral coefficients that capture different levels of spectral envelope information. Lower-order coefficients (e.g., C1) represent broad spectral characteristics, while higher-order coefficients (e.g., C2–C6 and beyond) capture progressively finer-scale variations in the log-Mel spectral envelope.*

DCT decorrelates the coefficients and provides a compact representation of the spectral envelope, with each MFCC encoding a distinct spectral pattern spanning the entire frequency range. This independence simplifies statistical modeling and allows features to be treated separately in probabilistic or machine learning frameworks. The representations from waveform to MFCC can be seen in Figure 2.3.

The first few MFCCs capture the broad shape of the spectral envelope, encoding most of the information relevant for distinguishing phonemes and syllables, such as the positions of the first and second formants and the overall resonant structure of the vocal tract (Davis and Mermelstein 1980). Higher-order coefficients preserve finer spectral details, including narrowband resonances, harmonic structure, and subtle timbral cues, which are important for speaker-specific characteristics, emotion, or expressive qualities of the voice (Davis and Mermelstein 1980). For tasks primarily focused on speech content or phoneme recognition, retaining only the lower-order coefficients is often sufficient, enabling a compact representation while preserving the acoustic cues essential for decoding linguistic information. In contrast, tasks such as speaker identification, emotion detection, or detailed voice quality analysis benefit from including higher-order coefficients to capture these more nuanced spectral features.

2.2 Styles of Speech and Infant-Directed Speech

A speech style can be thought as a consistent pattern in which a speaker modifies acoustic and expressive behaviors according to the social or communicative context. Speaking

styles manifest across a wide range of everyday situations, including casual conversation, read or public speech, emotional expression, and intimate interaction (Shinozaki et al. 2009). Although these contexts differ substantially, they share the property that speakers adjust both acoustic and linguistic characteristics in response to situational demands.

This section first reviews general speaking style variation across common communicative scenarios, as detailed in Subsection 2.2.1, highlighting differences in acoustic and linguistic properties between read, conversational, and emotional speech. The following subsection, Subsection 2.2.2, focuses on IDS, a specialized style used by caregivers when addressing infants. Here, both acoustic and linguistic adaptations are described, along with their functional relevance for infant attention, speech segmentation, and early language acquisition.

2.2.1 Read and Spontaneous Speech

Read speech is typically produced with deliberate planning and careful control of articulation. Speakers generally maintain a steadier pitch, slower and more predictable rhythm, and clearer realization of vowels and consonants (Dellwo et al. 2015; Lee and Kreiman 2022; Silverman et al. 1992). The linguistic structures in read speech often mirror written language, with utterances consisting of grammatically complete sentences and interruptions occurring mainly at sentence boundaries. When reading, speakers process the text in advance, using syntactic cues such as punctuation to regulate rhythm. Consequently, the acoustic signal is relatively consistent unless the reading is performed expressively, in which case some spontaneous elements may appear.

Spontaneous speech, in contrast, differs in multiple respects from its read counterpart. Human conversation typically involves interaction with another participant, introducing dynamics that are difficult to reproduce in read speech. Empirical studies suggest that spontaneous speech often exhibits generally rising prosody, while read speech typically shows declining patterns (Silverman et al. 1992). Prosody regarding F0 variability is also lower in spontaneous speech than in read speech (Dellwo et al. 2015). Utterances may be incomplete, and boundaries between them are less strictly defined. Spontaneous speech frequently includes interruptions, overlaps, hesitations, and repairs, all arising from the real-time nature of communication.

Another distinction stems from the recording environment. Natural conversations usually occur in uncontrolled settings, introducing noise and other acoustic variability. Spontaneous speech remains challenging to capture even in soundproof rooms with acted or improvised dialogue, as phenomena like overlapping speech persist. Artificially constructed scenarios may further reduce spontaneity because speakers are aware of the context. By contrast, read speech is often recorded in studios with professional equipment, producing clean, consistent audio. These differences demonstrate that, although both are forms of

speech, read and spontaneous speech pose fundamentally different modeling challenges and require distinct methodological approaches (Silverman et al. 1992).

2.2.2 Infant-Directed Speech

IDS, also historically referred to as “motherese” and sometimes as child-directed speech (CDS) when directed to slightly older children, is a distinctive communicative register used by caregivers when interacting with infants and young children (Bergelson et al. 2023; Snow 1972; Soderstrom 2007). IDS may play an important role in early language acquisition, as infants consistently appear to prefer IDS over adult-directed speech (ADS), and exposure to IDS has been associated with enhanced attention to speech and more effective early word learning (Fernald 1985; Frank et al. 2020; Huebner and Willits 2021; Soderstrom 2007; Song et al. 2010; Zhou et al. 2024). A related preference has also been observed for infant-directed singing compared to adult-directed singing (Masataka 1999).

IDS is characterized by systematic modifications in both acoustic and linguistic properties compared to adult-directed speech (ADS), which refers to speech that does not exhibit these modifications and is not directed to an infant or child. Thus, the distinction between IDS and ADS is context-dependent, as it relies on the communicative situation and the intended listener rather than purely acoustic cues. Caregivers typically use IDS extensively from birth through the first year of life, with a gradual transition toward more adult-like speech as children mature, often beginning around 6 to 12 months of age (Farran et al. 2016; Huebner and Willits 2021). Nevertheless, some features of IDS remain observable well into early childhood, in some cases up to around four years of age (Räsänen and Kocharov 2025). While IDS exhibits broadly consistent tendencies across contexts, its realization varies depending on the child’s age, the interactional setting, and individual caregiver differences. For instance, speech addressed to a three-month-old infant may differ substantially from speech directed to a three-year-old child, although both are categorized as IDS or CDS.

Because acoustic characteristics vary across contexts, speakers, and the age of the child, it is not straightforward to determine whether a given speech sample is truly child-directed based on acoustics alone. In synthetic speech, this contextual grounding is absent, making it difficult to assess whether the generated output successfully implements an appropriate IDS style. In particular, acoustic classification of IDS typically requires comparison with adult-directed speech (ADS) from the same speaker, as separating speaker-specific characteristics from style-related features is non-trivial (Williams and King 2019). This issue is further discussed in Subsection 2.5.1.

Acoustically, IDS is characterized by a higher mean F0, exaggerated intonation contours, slower speech rate, and increased pitch variability compared to ADS (Grieser and Kuhl

1988; Martin et al. 2016; Narayan and McDermott 2016; Soderstrom 2007). Some studies have also reported an expanded vowel space in IDS, suggesting clearer articulation of vowel categories (Cox et al. 2023; Kuhl et al. 1997; Odijk and Gillis 2023). However, other work has found null or inconsistent effects, indicating that this feature is not yet robustly established (Englund and Behne 2006; Martin et al. 2016; Miyazawa et al. 2017). As such, the precise acoustic characterization of IDS remains an open empirical question.

Linguistically, IDS tends to exhibit simplified vocabulary, shorter utterances, and reduced syntactic complexity compared to ADS (Cychosz, Edwards et al. 2021; Snow 1972; Soderstrom 2007). Repetition of words and phrases is common, and prosodic exaggeration is often used to highlight salient linguistic units. These features are thought to support early language acquisition by increasing infant engagement, enhancing attention, and improving the perceptual salience of linguistic input.

It has been found that IDS and CDS show largely stable acoustic tendencies across languages and culture (Cox et al. 2023), although the amount of IDS used varies somewhat across cultures; for example, Lebanese mothers have been observed to use IDS more frequently than American mothers (Farran et al. 2016). Regarding the influence of CDS, culture, and language on language learning, Bergelson et al. investigated the language output of 1,001 children aged 2 to 4 years across six cultures alongside the adult speech they heard, finding that greater exposure to adult speech at this age was associated with increased language production in children, indirectly suggesting that CDS input influences children's language output across cultures (Bergelson et al. 2023).

Because IDS is inherently tied to natural caregiver–child interaction, it is typically recorded in infants' everyday environments. Large-scale naturalistic corpora include the Providence Corpus (Demuth et al. 2006) and other datasets within the CHILDES database (MacWhinney and Snow 1985), which are widely used in child language acquisition research. The SAYCam corpus (Sullivan et al. 2021) provides longitudinal multimodal recordings of infants aged 6–32 months using head-mounted cameras, capturing both auditory and visual input over extended periods. The KIDS corpus (Zahner et al. 2016) contains annotated German IDS directed to children under one year of age.

In addition to naturalistic recordings, more controlled-environment datasets also exist. The ManyBabies 1 (MB 1) corpus (Frank et al. 2020) is an example of such a collection, in which caregivers were prompted to produce infant-directed speech in response to predefined objects. While these controlled settings offer higher recording quality and stronger experimental control, they come with their own limitations. Laboratory-based recordings typically yield smaller datasets than naturalistic daylong recordings, which can capture all speech input a child is exposed to over extended periods, rather than a limited number of hours recorded in a controlled setting. Additionally, in MB 1, the communicative context was restricted to object-centered descriptions (Frank et al. 2020). This constraint

can be considered advantageous for studies focusing on infants' attention to familiar versus unfamiliar objects in IDS. However, as noted by the authors, the structured elicitation—centering interactions around concrete objects—may have influenced the acoustic properties of the recorded infant-directed speech IDS (Frank et al. 2020).

2.3 Non-Neural Approaches to Text-to-Speech

TTS systems synthesize speech from written text input (Taylor 2009, pp.1–7). This involves mapping discrete, symbolic linguistic representations to continuous acoustic signals. The process includes converting graphemes or phonemes into speech sounds, predicting prosodic features, and synthesizing them into a continuous waveform. Effective TTS must preserve linguistic content while also conveying paralinguistic information, generating speech that is perceived as natural and consistent with human speech (Taylor 2009, pp. 26–34).

Before the widespread adoption of neural networks, TTS systems were primarily based on explicitly designed parametric models and other data-driven techniques. These approaches highlight key challenges in speech synthesis and form the basis of modern architectures. Over time, the field shifted from explicitly modeling human speech production to generating speech directly from data using machine learning methods.

This section roughly follows Paul Taylor's overview of TTS evolution (Taylor 2009), tracing the development from early source–filter methods to modern non-neural approaches to TTS. Classical source–filter synthesis (Subsection 2.3.1) includes articulatory, formant, and linear predictive synthesis. Concatenative and statistical parametric synthesis (Subsection 2.3.2) represent the main non-neural approaches that preceded neural TTS.

2.3.1 Classical source–filter synthesis approaches

Classical speech synthesis methods are largely based on the source–filter model, in which an excitation signal (source) is shaped by a vocal-tract filter to produce speech (Taylor 2009, pp. 360–361). The excitation is typically modeled as a periodic signal for voiced sounds and noise for unvoiced sounds. The main approaches differ in how the vocal tract is represented.

Formant-based synthesis is a source-–filter modeling technique that represents the vocal tract parametrically (Taylor 2009, pp. 398–399). Text is first converted into a sequence of phonemes using context-sensitive rules, and each phoneme is assigned a set of target acoustic parameters, including formant frequencies, bandwidths, excitation characteristics, pitch, and amplitude envelopes (Taylor 2009, pp.406–407). These parameters define the spectral envelope of the speech over time, with smooth transitions between phonemes achieved through interpolation or predefined rules.

The vocal tract's resonances are implemented using filters. For a single formant with frequency $\hat{F}$ and bandwidth $\hat{B}$, the transfer function can be written as: (Taylor 2009, pp. 399–400)

$$H_n(z) = \frac{1}{1 - 2e^{-2\hat{B}} \cos(2\pi\hat{F})z^{-1} + e^{-2\hat{B}}z^{-2}} \tag{2.8}$$

where $\hat{F}$ and $\hat{B}$ denote formant frequency and bandwidth.

Linear prediction (LP) synthesis simplifies formant-based synthesis by modeling the vocal tract as a single all-pole filter whose coefficients are automatically estimated from recorded speech using linear prediction analysis methods such as autocorrelation-based LP estimation (Bäckström et al. 2022, Ch. 3). In contrast to formant synthesis, where the vocal-tract transfer function is represented as a cascade of individual resonant filters corresponding to separate formants (Taylor 2009, pp. 410–411), LP synthesis models the vocal tract as a unified acoustic system with a single transfer function:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} \tag{2.9}$$

where the coefficients a_k approximate the resonant characteristics of the vocal tract.

Articulatory synthesis models speech production by simulating the physiological and aerodynamic behavior of the vocal tract, including structures such as the tongue, lips, jaw, and vocal folds (Taylor 2009, pp.416–418). The vocal tract is approximated as a sequence of time-varying acoustic tubes with different cross-sectional areas, through which airflow and sound propagation are computed (see Figure 2.4).

Although source-–filter models can produce intelligible speech, their simplified acoustic representation of speech production can result in artificial-sounding output. Articulatory synthesis can generate more realistic speech by modeling the physical speech-production mechanism directly, but its anatomical detail and computational complexity make it difficult to generalize across speakers (Tom et al. 2001).

2.3.2 Concatenative and Statistical parametric synthesis

Concatenative synthesis was the dominant approach prior to statistical parametric methods and does not follow the source-–filter framework. It generates speech by selecting and concatenating short prerecorded units, typically diphones, which encode transitions between phonemes (Bäckström et al. 2022, Ch. 9; Taylor 2009, pp.425–426). These units are stored in a large, annotated database, allowing synthesis to avoid explicit acoustic modeling or parameter generation.

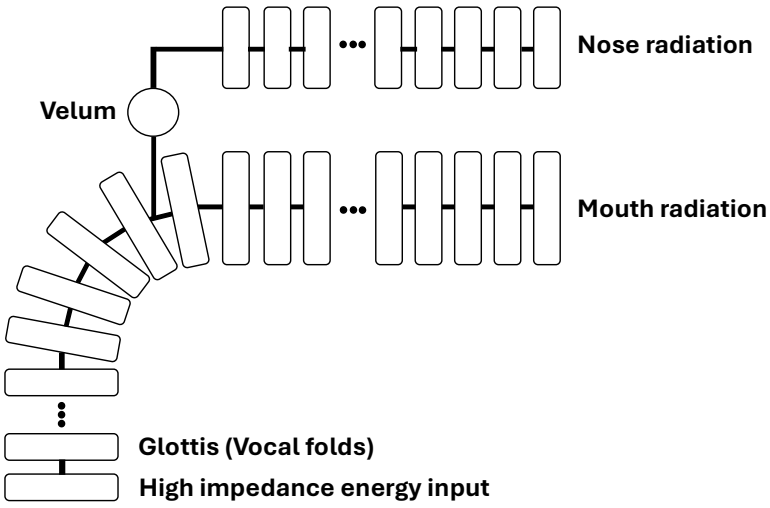

Figure 2.4. *Simplified tube model of the human vocal tract for articulatory speech synthesis, in which airflow from a glottal source excites acoustic wave propagation through area-varying tube sections (white rectangles), producing resonances arising from reflections, cross-sectional changes, and radiation losses. Oral and nasal branches are coupled via the velum, with separate radiation at the lips and nostrils. Reproduced from (Hill et al. 1995)*

Direct use of recorded waveforms can yield natural-sounding speech when there is sufficient coverage of speech contexts, as the characteristics of the original voice are preserved. However, achieving comprehensive coverage requires extensive, carefully curated corpora, making database construction both time-consuming and costly. The TTS system is also inherently constrained in terms of speakers, as it can only produce voices that are present in the recorded database. Additionally, concatenation points—particularly at word boundaries—can introduce audible discontinuities or unnatural transitions (Bäckström et al. 2022, Ch. 9; Taylor 2009, pp. 443–445).

Statistical parametric synthesis was the dominant non-neural approach to TTS prior to deep learning methods. It models speech as a sequence of hidden probabilistic states and generates acoustic features using statistical sequence models, most commonly hidden Markov models (HMMs) combined with decoding algorithms such as Viterbi (Bäckström et al. 2022, Ch. 9; Rabiner 1989; Tokuda et al. 2000). In this framework, speech is represented as a progression through hidden states, each associated with a probability distribution over acoustic features such as MFCC (Bäckström et al. 2022, Ch. 9).

In the context of TTS, an HMM parameterized by λ can be defined as

$$\lambda = (A, B, \pi), \tag{2.10}$$

where A encodes transition probabilities between phonetic or sub-phonetic states, B defines state-dependent observation distributions over acoustic features, and π specifies the initial state distribution (Rabiner 1989; Tokuda et al. 2000). Given a linguistic input, synthesis is performed by inferring the most likely state sequence using the search algorithm. This produces a sequence of state-aligned acoustic parameters representing the predicted spectral and excitation characteristics of speech over time (Bäckström et al. 2022, Ch. 9).

These resulting parameter trajectories are then converted into a waveform using a vocoder, which reconstructs speech from acoustic features such as the spectral envelope, fundamental frequency, and aperiodicity (Bäckström et al. 2022, Ch. 9). For example, the Griffin–Lim (Griffin and Lim 1984) algorithm performs iterative phase reconstruction from a target magnitude spectrogram, alternating between time and frequency domains until the synthesized waveform becomes consistent with the desired spectral magnitude.

Limitations of HMM-based TTS arise from representing continuous speech with discrete hidden states under restrictive probabilistic assumptions. Acoustic features are typically modeled using Gaussian or Gaussian mixture distributions, which are limited in capturing the highly non-linear and multi-modal nature of real speech dynamics. Temporal dependencies are restricted to short-term state conditioning rather than long-range context. State representations capture only averaged acoustic properties, limiting fine-grained temporal modeling and often causing discontinuities at state transitions (Taylor 2009, pp.483).

2.4 Neural Approaches to Text-to-Speech Synthesis

Neural TTS represents a shift toward fully data-driven synthesis, building on concepts from statistical parametric methods. Unlike HMM-based TTS, neural approaches learn pronunciation, timing, and acoustic variation directly from paired text–speech data within a fully differentiable framework, relying on training data rather than manually specified rules. This transition enabled substantial improvements in naturalness, flexibility, and overall speech quality compared to previous non-neural methods (Shen et al. 2018; Tachibana et al. 2018; Zen et al. 2013).

Most neural TTS systems decompose speech synthesis into two stages: acoustic modeling and vocoding. The acoustic model predicts an intermediate representation from text, typically a mel-spectrogram, while a vocoder converts this representation into a waveform. Both stages have evolved from traditional signal-processing approaches toward learned neural models (Oord, Dieleman et al. 2016; Oord, Li et al. 2017).

At a high level, neural acoustic models map linguistic input, such as text or phoneme sequences, to acoustic representations. Early neural TTS architectures (Wang, Skerry-

Ryan et al. 2017) generated acoustic features sequentially, predicting each output frame conditioned on previous predictions. While effective, this autoregressive formulation limited computational efficiency and made long-sequence modeling more challenging. Later architectures introduced more parallel processing strategies (Kim et al. 2020; Ren et al. 2019), enabling faster synthesis while maintaining high speech quality. Over time, improvements in architectural design progressively enhanced the ability of TTS models to capture both short- and long-term dependencies in speech data, approaching the performance of modern systems.

Recent TTS systems have further extended these models by enabling zero-shot and cross-speaker synthesis through conditioning on speaker embeddings or reference audio (Casanova, Davis et al. 2024). Speaker and style control in TTS is further discussed in Section 2.5, while this section focuses on general TTS architecture. In this section, neural TTS architectures and key design considerations from early to modern systems are discussed. Sequence-to-sequence models with early attention mechanisms are first presented in Subsection 2.4.1. The role of convolutional networks in improving local context representation for neural TTS is described in Subsection 2.4.2. Modern transformer-based architectures, which employ more complex attention mechanisms to enable efficient modeling of long-range dependencies across entire sequences, are examined in Subsection 2.4.3. Lastly, neural vocoders are covered in Subsection 2.4.4.

2.4.1 Sequence-to-Sequence Architectures

Sequence-to-sequence (seq2seq) models with attention were the first major end-to-end neural approach to text-to-spectrogram synthesis (Sutskever et al. 2014; Zen et al. 2013). These models learn a conditional mapping from an input sequence $x = \{x_1, \ldots, x_N\}$ (text tokens such as characters or phonemes) to an output acoustic sequence $y = \{y_1, \ldots, y_T\}$, typically modeled as (Sutskever et al. 2014):

$$p(y \mid x) = \prod_{t=1}^{T} p(y_t \mid y_{<t}, x) \tag{2.11}$$

This autoregressive formulation means that each predicted acoustic frame depends on both the input text representation and all previously generated output frames. The input sequence is first encoded into a sequence of hidden representations:

$$h_n^{enc} = \text{Encoder}(x_n, h_{n-1}^{enc}), \tag{2.12}$$

where h_n^{enc} captures linguistic and contextual information from the input sequence.

To align variable-length text inputs with acoustic outputs, seq2seq TTS models introduced

attention mechanisms (Bahdanau et al. 2015). At each decoding step, attention computes a context vector:

$$c_t = \sum_{n=1}^{N} \alpha_{t,n} h_n^{enc}, \tag{2.13}$$

where the attention weights are normalized over encoder states:

$$\alpha_{t,n} = \frac{\exp(e_{t,n})}{\sum_{k=1}^{N} \exp(e_{t,k})}, \tag{2.14}$$

and the alignment score

$$e_{t,n} = \mathbf{v}^\top \tanh(W_s s_{t-1} + W_h h_n^{enc}) \tag{2.15}$$

measures how relevant encoder state h_n^{enc} is for generating the current output frame.

The decoder then generates acoustic frames sequentially:

$$y_t = \text{Decoder}(y_{t-1}, c_t, s_{t-1}), \tag{2.16}$$

where generation is conditioned on the previously predicted frame, the attention context, and the decoder state.

Early seq2seq TTS systems commonly used recurrent neural networks (Wang, Skerry-Ryan et al. 2017). In particular, Long Short-Term Memory (LSTM) architectures (Greff et al. 2017) were used because they were effective at modeling temporal dependencies across speech sequences.

In a typical TTS pipeline, input text tokens are first converted into learned embeddings and processed by the encoder, often implemented using bidirectional recurrent layers. The decoder then predicts mel-spectrogram frames autoregressively, while attention provides soft alignment between text positions and acoustic frames.

One of the earliest influential seq2seq TTS models, Tacotron (Wang, Skerry-Ryan et al. 2017), follows this architecture: text embeddings $\rightarrow$ encoder $\rightarrow$ attention $\rightarrow$ decoder. The predicted mel-spectrogram is subsequently converted into a waveform using a vocoder.

2.4.2 Convolutional Neural Networks in Sequence-to-Sequence TTS

Convolution is a mathematical operation that combines an input sequence with a small set of learnable weights (a kernel) to produce a new representation. At each position, the kernel is applied to a local neighborhood of the input, allowing the model to extract local patterns (Goodfellow et al. 2016) [pp. 326–327].

A discrete 1D convolution over a sequence can be written as (LeCun et al. 1998):

$$y_i = \sum_{k=0}^{K-1} w_k \, x_{i+k}, \tag{2.17}$$

where x_i is the input sequence, w_k are learnable kernel weights, and y_i is the output at position i.

Convolutional neural networks (CNNs) were introduced into sequence-to-sequence TTS models to improve the representation of local contextual structure in speech, as spectrograms provide a two-dimensional time–frequency representation in which convolutional kernels can effectively extract local patterns along both axes. In this setting, stacking convolutional layers increases the receptive field, enabling the modeling of increasingly hierarchical dependencies, while architectural variants such as dilated convolutions and residual connections further improve contextual coverage and training stability (He et al. 2016; Yu and Koltun 2016). Strided convolutions are additionally used to reduce temporal resolution and produce more compact representations (Goodfellow et al. 2016, pp. 343–344).

In early systems such as Tacotron 2 (Shen et al. 2018), convolutional layers were introduced in the encoder to process embedded text sequences, replacing the need for purely recurrent feature extraction at the input stage. These layers jointly process neighboring tokens to capture local patterns such as short phoneme sequences and coarticulation effects, while a subsequent bi-directional recurrent encoder models longer-range dependencies (Wang, Skerry-Ryan et al. 2017). Later models, such as Deep Voice 3 (Ping et al. 2018), extended this idea to fully convolutional architectures, replacing recurrent layers with stacked convolutional blocks with residual connections in both encoder and decoder, improving parallelism.

2.4.3 Transformers and Transformer-Based Architectures

Transformer architectures represent the most recent advancement in modeling long-range dependencies in text-to-speech synthesis. They extend previous approaches by using self-attention, which computes relationships between all positions in a sequence simultaneously (Vaswani et al. 2017). Unlike recurrent and convolutional architectures, where dependencies are propagated through sequential or local operations, self-attention enables direct global interactions between all tokens in a single layer.

Internally, a Transformer layer consists of multi-head self-attention and a position-wise feed-forward network (Vaswani et al. 2017). Given an input sequence $X = [x_1, \ldots, x_N]$, the queries, keys, and values are computed as linear projections (Vaswani et al. 2017):

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V, \tag{2.18}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are learnable matrices.

Scaled dot-product attention is defined as (Vaswani et al. 2017):

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V, \tag{2.19}$$

where the softmax function, which normalizes the output values into a probability distribution, is applied row-wise over the keys for each query position.

Multi-head attention applies h parallel attention heads, allowing the model to attend to different types of relationships between sequence elements simultaneously (Vaswani et al. 2017):

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{2.20}$$

where the outputs of all attention heads are concatenated and linearly projected using the learnable output matrix W^O. Each attention head is defined as (Vaswani et al. 2017):

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{2.21}$$

where W_i^Q, W_i^K, and W_i^V are learnable projection matrices that map the input queries, keys, and values into head-specific representations.

The feed-forward network is applied independently at each sequence position (Vaswani et al. 2017):

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2, \tag{2.22}$$

where $W_1 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$ and $W_2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$ are learnable weight matrices that project the input from the model dimension d_{model} to the higher-dimensional feed-forward space d_{ff} and back. The ReLU activation introduces non-linearity between the two linear transformations.

Each Transformer sublayer is wrapped with a residual connection followed by layer normalization (Vaswani et al. 2017):

$$\text{LayerOutput} = \text{LayerNorm}(x + \text{Sublayer}(x)), \tag{2.23}$$

where Sublayer$(\cdot)$ denotes either the multi-head attention operation or the feed-forward network defined in Equations 2.20 and 2.22, respectively. The residual connection, formed by adding the input x to the sublayer output Sublayer(x), helps preserve information across layers, while layer normalization improves training stability.

One of the first TTS models to apply Transformers was Transformer TTS (Li et al. 2019), which replaced recurrent layers in Tacotron2 (Shen et al. 2018) with self-attention modules. Later, FastSpeech (Ren et al. 2019) proposed a non-autoregressive Transformer-based architecture. Unlike autoregressive models, which implicitly determine output length through sequential generation, non-autoregressive models generate all outputs in parallel and therefore require an explicit mechanism for constructing frame-level alignment. In FastSpeech, this is achieved using a length regulator, which expands text tokens according to predicted durations to match the length of the acoustic frame sequence (Ren et al. 2019).

2.4.4 Neural Vocoders

In addition to spectrogram prediction models, vocoders gradually progressed toward neural architectures. As an example of a neural vocoder, WaveNet (Oord, Dieleman et al. 2016)is a neural model which generates audio waveforms from spectrograms. Its main component is a stack of dilated causal convolutional layers, where each layer convolves over a subset of past samples with increasing dilation (Yu and Koltun 2016). This enables the network to capture short-term waveform details as well as medium- and long-range temporal dependencies necessary for natural speech generation.

WaveNet employs residual and skip connections to facilitate learning and combine features from different temporal scales (He et al. 2016). Residual connections pass the input of a layer directly to a later layer, enabling the network to learn modifications to the input rather than the entire transformation. Skip connections aggregate outputs from multiple layers and combine them in the final output layers, allowing the network to integrate short-term waveform details with longer-term speech structures (Oord, Dieleman et al. 2016).

The network models the conditional distribution of each audio sample x_t given all previous samples $x_{<t}$ using a softmax over quantized audio values encoded with μ-law at 8-bit resolution (Oord, Dieleman et al. 2016):

$$p(x_t \mid x_{<t}) = \mathrm{Softmax}(z_t), \quad z_t = f(x_{<t}; \theta), \tag{2.24}$$

where $f(\cdot;\theta)$ denotes the WaveNet convolutional network and θ its learnable parameters. The vector z_t contains unnormalized logits over the discrete μ-law quantized audio value

space.

WaveNet can also be conditioned on external features, such as mel-spectrograms produced by a preceding text-to-spectrogram model (Shen et al. 2018). The conditional distribution becomes:

$$p(x_t \mid x_{<t}, h) = \mathrm{Softmax}(z_t), \quad z_t = f(x_{<t}, h; \theta), \tag{2.25}$$

where h denotes the conditioning acoustic features (e.g., a mel-spectrogram), and z_t represents the corresponding logit vector conditioned on both past samples and h.

where h represents the acoustic features. Conditioning can be applied either globally across the sequence or locally at each time step (Oord, Dieleman et al. 2016).

Later neural vocoders, including Parallel WaveGAN (Yamamoto et al. 2020) and HiFi-GAN (Kong et al. 2020), adopt fully convolutional generative adversarial architectures. In these models, the generator transforms an input spectrogram into a time-domain waveform segment, while the discriminator evaluates whether the generated waveform matches real speech. Unlike WaveNet, these GAN-based vocoders are non-autoregressive, producing all samples in a segment simultaneously in parallel, which significantly improves inference efficiency (Kong et al. 2020; Yamamoto et al. 2020).

2.5 Speaker and Style Control in Neural TTS

Controlling speaker identity and speech style is a central challenge in modern neural text-to-speech (TTS) synthesis. Early non-neural and initial neural models focused more primarily on generating intelligible speech from text, emphasizing accurate pronunciation and the transfer of information. As models evolved, it became increasingly possible to produce speech that is more human-like, both in speaker characteristics and in expressive style, making the control of these features a central focus. Ideally, a TTS system could replicate the characteristics of any speaker, including those not present in the training data, while dynamically applying stylistic features depending on the communicative context.

This section provides an overview of methods for representing and manipulating speaker identity and speech style in neural text-to-speech (TTS) systems. First, the concepts of speaker identity and style are discussed in Subsection 2.5.1. Next, early approaches to dynamic speaker modeling are reviewed in Subsection 2.5.2, focusing on multi-speaker embeddings. Systems such as Deep Voice 2 and Deep Voice 3 enabled control over voices present in the training data, but did not generalize to unseen speakers. Subsequently, Subsection 2.5.3 introduces more recent zero-shot TTS architectures, which extend speaker and style control to previously unseen speakers by conditioning synthesis

on short reference audio samples. Finally, Subsection 2.5.4 examines text-driven style control, where natural language prompts or structured tags embedded in the input text are used to guide expressive aspects of synthesized speech.

2.5.1 Speaker Identity and Speech Style

In speech processing, it is important to distinguish between speaker identity features and speech style features (see Kinnunen and Li 2010 for a review). Speaker identity features are intrinsic properties of an individual speaker that tend to remain relatively stable across utterances and contexts, whereas style features are more dependent on communicative situation and speaking context (Williams and King 2019). Speaker identity is shaped by anatomical differences in the vocal apparatus, which give rise to characteristic voice qualities, as well as by long-term speaking habits influenced by culture and environment.

In practice, no acoustic features are exclusively tied to either identity or style. Instead, speaker recognition systems rely on combinations of features capturing different aspects of speech production, including short-term spectral characteristics related to vocal tract resonances, voice source properties, prosodic patterns, and higher-level linguistic behavior, as reviewed by Kinnunen and Li (2010). These representations are optimized to increase between-speaker variability while reducing within-speaker variability, within which style-related variation is typically included.

Speaker verification is a sub-problem of speaker recognition in which a model addresses the binary classification task of determining whether two speech samples originate from the same speaker. A prominent example of such a model is ECAPA-TDNN (Desplanques et al. 2020). Recent research has also shown that modern verification systems can remain robust even when speakers attempt to disguise their identity (Jung et al. 2022).

From the verification perspective, one may consider an alternative interpretation: if speaker identity can be robustly modeled from speech, even when speakers modulate their voice, then aspects of speech style may also be amenable to modeling when treated as part of within-speaker variation (Williams and King 2019). This could motivate extending the verification paradigm to a “style verification” setting, in which the task is to determine whether two speech samples exhibit similar speaking style, rather than whether they originate from the same speaker.

2.5.2 Multi-Speaker and Style Embeddings

Early neural TTS models could only reproduce voices present in the training data (Arik et al. 2017; Ping et al. 2018; Shen et al. 2018). Multi-speaker TTS models, such as Deep Voice 2 (Arik et al. 2017) and Deep Voice 3 (Ping et al. 2018), assigned a learnable embedding vector to each training speaker. During synthesis, selecting a speaker

embedding produced speech in that speaker's voice. However, this approach could not generalize to speakers outside the training set.

Style control was introduced in models such as GST-Tacotron (Wang, Stanton et al. 2018), which used learnable style tokens and an attention module to generate a style vector capturing prosody, pitch movement, or expressiveness. These tokens allowed interpolation between learned styles but did not reliably separate speaker identity from style.

Transformer-based models, such as FastPitch (Lancucki 2021), introduced explicit pitch prediction and conditioning. Each mel-spectrogram frame receives a predicted pitch embedding derived from text via a duration predictor. This allows manipulation of intonation and style by modifying the predicted pitch contour.

2.5.3 Zero-Shot Speaker Adaptation

Modern zero-shot TTS models enable synthesis of previously unseen speakers using only a short reference audio clip (Betker 2023; Casanova, Davis et al. 2024; Wang, Chen et al. 2023). A reference encoder extracts a fixed-dimensional representation from an audio sample, summarizing both vocal characteristics and prosody. This representation conditions the acoustic model, guiding spectrogram generation to match the reference speaker's identity and style (Casanova, Davis et al. 2024; Skerry-Ryan et al. 2018; Wang, Stanton et al. 2018).

The reference representation can be integrated into the acoustic model by concatenation, addition to intermediate states, or modulation of attention and feed-forward layers (Casanova, Davis et al. 2024; Skerry-Ryan et al. 2018). Systems such as Tortoise TTS, xTTS-v2, and VALL-E (Betker 2023; Casanova, Weber et al. 2022; Wang, Chen et al. 2023) encode speaker identity and prosody, enabling reproduction of reference speech characteristics.

2.5.4 Text-Driven Style Control

In addition to reference audio, natural language prompts or tags can guide prosody, emotion, or speaking style in neural TTS (Shin et al. 2022; Yang et al. 2023). Text inputs may include descriptors like "angrily," "softly," or structured tokens encoding stylistic features. The TTS model conditions on these prompts to adjust pitch, duration, amplitude, and spectral features, producing speech that reflects the requested style without a reference sample.

Commercial systems, such as ElevenLabs' neural TTS models (ElevenLabs 2025b; ElevenLabs 2025c), combine multiple style control mechanisms and optional parameters for expressive speech. Models like `eleven_multilingual_v2`, `eleven_v3`, and `flash_v2.5`

analyze phonetic and broader text context to generate emotionally aware speech. Users can control stability, similarity to a reference voice, style, intonation, pacing, and expressive quality. Text-embedded style cues, such as `[whispering]` or `[laughs]`, can trigger localized vocal effects directly from the text (ElevenLabs 2025a).

2.6 Preprocessing Speech Data

Speech data exhibits variability in multiple ways. Some of this variability arises from internal factors, such as anatomical differences in speakers' vocal folds and articulatory configurations, which influence pitch, timbre, and phonation patterns. Other variability is external, resulting from environmental conditions, recording equipment, or background noise, meaning that not all information captured in the speech signal originates from the speech itself.

Disentangling speech from other acoustic components is inherently approximate, since relevant and irrelevant components often overlap in both time and frequency domains. Consequently, reducing noise can inadvertently affect the speech signal itself, creating a trade-off between signal purity and preservation of essential information. The required level of cleanliness also depends on whether the system aims to generate speech, as in TTS, or simply discriminate between classes, as in speech recognition. Preprocessing therefore aims to reduce task-irrelevant variability while preserving the aspects of the signal critical for the intended application.

Some speech processing systems do not require or even desire noise removal before processing. For instance, neural models trained for ASR often encounter noisy input at inference, and the system must learn to handle these conditions. In this subsection, preprocessing with respect to noise reduction refers specifically to any processing applied to the data prior to the task, rather than internal noise handling performed by the system.

The purpose of the section is to review common speech enhancement techniques for speech data with focus on the preprocessing needing to be done for speech data used in conjunction of TTS systems. The Subsection 2.6.1 focuses on noise reduction and audio enhancement methods that improve signal quality and reduce unwanted variability. Subsection 2.6.2 discusses how data quality, recording conditions, and preprocessing choices impact the training and inference of neural text-to-speech models, highlighting the importance of well-prepared datasets for high-fidelity synthesis.

2.6.1 Noise Reduction and Audio Enhancement

A central concern in speech processing is external noise, which refers to acoustic energy unrelated to the target speech signal, including background noise, reverberation, echoes, microphone artifacts, or overlapping speakers (Bäckström et al. 2022, Ch.11). This sub-

section focuses on denoising and neural enhancement methods.

Classical approaches include spectral subtraction and Wiener filtering. Spectral subtraction estimates the noise spectrum during non-speech segments and subtracts it from the noisy signal (Bäckström et al. 2022, Ch. 11; Boll 1979). In the short-time Fourier transform (STFT) domain, the magnitude estimate is given by (Boll 1979):

$$|\hat{S}(t,f)| = \max\big(|Y(t,f)| - |N(t,f)|, 0\big), \tag{2.26}$$

where $Y(t,f)$ is the noisy STFT and $N(t,f)$ is the estimated noise magnitude spectrum. The complex-valued estimate is reconstructed using the noisy phase:

$$\hat{S}(t,f) = |\hat{S}(t,f)|e^{j\angle Y(t,f)}. \tag{2.27}$$

Wiener filtering minimizes mean-square error under stationarity assumptions. The Wiener gain is (Ephraim and Malah 1984):

$$G(t,f) = \frac{P_S(t,f)}{P_S(t,f) + P_N(t,f)}, \tag{2.28}$$

and the enhanced signal is:

$$\hat{S}(t,f) = G(t,f)\, Y(t,f). \tag{2.29}$$

While classical methods assume relatively stationary noise, neural speech enhancement models learn a data-driven mapping from noisy to clean speech. Neural enhancement systems are trained on paired noisy and clean speech to learn a mapping f_θ:

$$\hat{S} = f_\theta(Y), \tag{2.30}$$

where $\hat{S}$ is the predicted clean speech and f_θ is typically parameterized by deep neural networks.

Modern systems such as Resemble Enhance decompose this process into denoising followed by perceptual enhancement (ResembleAI 2023).

First, a denoising network estimates a mask:

$$M = g_{\theta_1}(Y), \tag{2.31}$$

which is applied to the input spectrogram:

$$\hat{S}_{\text{denoise}} = M \odot Y, \tag{2.32}$$

where g_{θ_1} is a neural network and M is a learned time–frequency mask.

The denoised representation is then refined in a learned latent space, i.e., a feature space produced by a neural network in which the representation is no longer directly interpretable as a waveform or spectrogram but encodes higher-level acoustic structure:

$$Z = f_{\theta_2}(\hat{S}_{\text{denoise}}), \tag{2.33}$$

where Z denotes a latent acoustic representation.

Finally, a generative model and vocoder reconstruct the enhanced waveform:

$$\hat{S}_{\text{enhanced}} = \text{Vocoder}(Z), \tag{2.34}$$

While neural methods are highly effective for non-stationary and overlapping noise, overly aggressive denoising can remove parts of the speech signal and distort its spectral structure, reducing naturalness and intelligibility. Additionally, in perceptual metric-driven neural enhancement systems, hallucination-like artifacts may occur, where the model generates speech components not present in the original signal, particularly under severe degradation. These hallucinations arise because the model tries to minimize its loss function under degraded input conditions, where noise limits the amount of speech information that can be reliably reconstructed (Close et al. 2024).

2.6.2 Data Quality and Quantity Requirements for Neural TTS Systems

Neural TTS models are generally sensitive to suboptimal audio quality in training or reference data, as suggested by recent work that explicitly addresses noisy or non-ideal training conditions (Lu et al. 2025; Zhu et al. 2023). Acoustic features from speech and non-speech components often occupy overlapping frequency ranges, making it non-trivial for a model to determine which parts of the signal correspond to speech. If issues with the audio quality are not addressed, acoustic features may be distorted and incorporated into the learned speech representation (Lu et al. 2025; Zhu et al. 2023). This makes it challenging for a TTS model to produce intelligible and natural-sounding speech when trained on audio containing noise, recording artifacts, or other irrelevant acoustic content. As a result, it has been common practice to train TTS models on originally clean or denoised data (Zhang, Ren, Tan et al. 2020), rather than directly learning from noisy or unfiltered

recordings.

In contrast, some neural systems, such as ASR, are designed to operate under non-ideal conditions and tolerate noise. Tasks such as ASR or speaker recognition are typically formulated as discriminative classification problems rather than generative modeling tasks. Consequently, the objective is to predict discrete labels, such as phonemes, words, or speaker identities, rather than reconstruct the full acoustic signal. This allows the models to learn representations that emphasize task-relevant information while suppressing irrelevant acoustic variation, making them generally more robust to noise and signal degradation.

In addition to the requirement for clean data, modern neural TTS systems—and large neural models more generally—also require substantial amounts of training data. Single-speaker TTS models typically require tens of hours of high-quality recordings to reliably learn phonetic, prosodic, and acoustic patterns (Ping et al. 2018; Wang, Stanton et al. 2018). The increasing scale of modern neural architectures is reflected in both model size and training data requirements across generations of TTS systems. For example, Tacotron (Wang, Skerry-Ryan et al. 2017) contains on the order of tens of millions of parameters and was trained on roughly 24 hours of single-speaker speech, whereas more recent systems such as Tortoise integrate large autoregressive language modeling components based on GPT-2–style architectures (Radford, Wu et al. 2019), with parameter counts in the hundreds of millions range depending on configuration, and are trained on thousands of hours of multi-speaker and multi-style speech data.

As a note, the fact that neural models have many parameters or require large amounts of training data is not inherently “good” or “bad”, although larger models require more computational resources to train. Rather, the data and complexity requirements of neural TTS models increase substantially when scaling to multilingual, multi-style, or multi-speaker settings due to the combinatorial growth in phonetic, prosodic, and contextual patterns that must be represented.

The requirement for large amounts of clean data is one reason why directly modeling low-resource speech with neural TTS is challenging. Low-resource settings arise when only limited amounts of speech data are available for a given speaker, style, or language, often due to the difficulty or cost of collecting recordings under consistent conditions. This is particularly problematic when the target speaker or speaking style is difficult to capture in sufficient quantity without constraining naturalness or expressiveness, such as in the case of capturing high-quality IDS. As a result, training neural TTS models that can reliably generate high-quality speech from limited or low-quality data remains a significant challenge.

2.7 Technological Applications of IDS Modeling

IDS plays a role in early language acquisition by capturing infants' attention, enhancing engagement, and facilitating the learning of speech sounds and linguistic structures, as discussed in Subsection 2.2.2. Reliable IDS generation could provide new opportunities for computational developmental language research and speech-based assistive technologies. Advances in IDS-capable TTS systems would enable practical applications ranging from social robotics to interactive learning tools. At the same time, IDS modeling presents several challenges, including capturing natural variability, modeling developmental changes across age groups, and ensuring intelligible and natural-sounding output in spontaneous or noisy conditions. Addressing these challenges is essential for designing systems that effectively support learning, engagement, and social interaction in children.

Figure 2.5. *Pepper the robot in a mall, 27 August 2016. The photograph depicts a child interacting with the social robot Pepper, illustrating potential applications of IDS synthesis in robotics and early childhood learning contexts. Photo by Nesnad, via Wikimedia Commons, licensed under CC BY 4.0.*

One application that could benefit from an IDS-capable TTS system is social robotics. Research on social robotics in educational contexts spans primary school classrooms

(Ahtinen and Kaipainen 2020), early childhood and kindergarten gamified activities (Samposalo et al. 2024), in-home family co-learning (Ahtinen, Beheshtian et al. 2023), collaborative learning events such as Robocamp and Robotour (Ahtinen, Chowdhury et al. 2023), and university-level programming workshops (Quintero and Ahtinen 2023). While most studies focus on older children, in Finland children can enter early childhood education from the month they turn nine months (*Act on Early Childhood Education and Care (540/2018)* 2018). A social robot capable of producing IDS could therefore facilitate early interactions with children at this developmental stage, supporting both learning and engagement.

Other interactive tools, applications, and social facilitators designed to support early language acquisition could also benefit from IDS-capable TTS. Infants are more likely to focus on an object that an adult is looking at and describing using IDS than the same object described with adult-directed speech, highlighting the importance of IDS for attention and early learning (Fernald 1985; Zhou et al. 2024).

Socially responsive systems, in which speech adapts contingently to a child's actions or attention, are more effective than passive or pre-recorded audio (Roseberry et al. 2014). These kinds of systems would benefit from models capable of real-time interaction. Many current apps or toys claim to be interactive but rely on pre-recorded responses, which lack the adaptive benefits of truly responsive IDS-driven interaction that mirrors live social engagement. Such dynamic interaction may be particularly important for children under four, when attentional cues and contingent input strongly influence language acquisition (Roseberry et al. 2014).

As a specific example, artificial IDS delivered through interactive objects or systems may support autistic children, who allocate attention differently to social versus non-social stimuli, often focusing more on objects than on people (Chawarska et al. 2010). From a monotropism perspective, attention is distributed more narrowly toward areas of interest, limiting overall attentional scope (Murray 2018). Some studies suggest that IDS may not be as intrinsically engaging for autistic children as it is for neurotypical peers (Droucker et al. 2013), possibly reflecting differences in social attention and processing. Embedding IDS within objects or responsive systems could therefore align with these attentional preferences by situating communicative input within objects of interest, scaffolding attention, language processing, and the learning of social conventions in ways that require less reliance on sustained joint attention to a human interlocutor.

More broadly, in childcare contexts, a responsive IDS system could supplement caregiver interaction for all children, providing additional opportunities for engagement, particularly in settings such as daycare centers where one adult supervises multiple children and has limited capacity for individualized interaction.

3. METHODS

Given the limited size and variable quality of available infant-directed speech (IDS) datasets (Subsection 2.2.2) and the limitations of neural TTS models in terms of training data quantity and quality required (Subsection 2.6.2), training a dedicated IDS-specific TTS system from scratch was not feasible. To address these limitations, this study applies speech enhancement techniques to improve the quality of natural IDS recordings and leverages an existing zero-shot reference-encoder TTS model to transfer IDS-specific acoustic features from reference audio to synthesized speech.

This chapter describes the methodology used to implement the text-to-IDS synthesis pipeline and the experimental design for its evaluation. The central objective is to evaluate whether natural, potentially enhanced IDS reference recordings can be used to generate intelligible synthetic IDS while preserving key acoustic and prosodic characteristics of the speaking style. This objective is addressed through four research questions:

1. How effectively does the TTS model transfer IDS-specific stylistic attributes from natural IDS reference recordings to synthesized speech?
2. How intelligible is the synthesized speech when natural IDS recordings are used as reference inputs?
3. How does denoising and enhancement of IDS reference audio affect synthesis quality and style transfer?
4. How do IDS-specific linguistic characteristics influence synthesized prosody and intelligibility?

Section 3.1 describes the speech corpora and the collection of text prompts used in the study, as well as the neural TTS and enhancement models in the text-to-IDS synthesis pipeline. Section 3.2 presents the acoustic and intelligibility analysis methods used in the study. Section 3.3 defines the experimental design used to assess prosody transfer and intelligibility across synthesis conditions.

3.1 Data and the Text-to-Speech pipeline

This section describes the data and text-to-IDS pipeline used in the experiment. Subsection 3.1.1 presents the ManyBabies 1 (MB 1) corpus (Frank et al. 2020) and the subset of

it used in this study. Subsection 3.1.2 presents the TIMIT corpus (Garofolo et al. 1993). Subsection 3.1.3 describes the models and methods used for textual prompt generation for text-to-speech (TTS) purposes. Finally, the neural models used in the TTS pipeline are introduced: the neural enhancer used for reference preprocessing in Subsection 3.1.4, and the xTTS-v2 model (Casanova, Davis et al. 2024) in Subsection 3.1.5.

3.1.1 ManyBabies 1 Corpus

The IDS reference audio used in this study was derived from the MB 1 corpus (Frank et al. 2020). The corpus contains semi-controlled natural recordings of IDS and adult-directed speech (ADS) produced by caregivers. In the recordings, mothers speak either to their infants (IDS condition) or to an adult experimenter (ADS condition), allowing comparisons of speaking style within the same speakers.

For the MB 1 study, the North American English (NAE) recordings were obtained from 15 mothers of NAE-learning infants. Speech was recorded in two laboratories, one in Canada and one in the United States. The recording procedures were consistent across both sites, providing an infant-friendly environment and using lapel microphones connected to smartphone recording applications to capture the mothers' speech (Frank et al. 2020). During the recordings, caregivers were asked to describe, one at a time, a set of familiar and unfamiliar objects presented to the infant or the experimenter.

For the experiment in this study, a subset of 150 utterances was extracted from the corpus, consisting of 75 IDS and 75 ADS utterances. The subset was constructed to ensure comparability across speakers, meaning that an equal number of natural utterances was selected from each speaker. Additionally, it was preferable to include as much speech duration as possible from each utterance, while maintaining minimal discrepancy between the IDS and ADS conditions in terms of per-speaker duration.

It is reported in Frank et al. (2020) that, in the constructed MB 1 audio stimuli used in their study, IDS utterances are on average shorter than ADS utterances. A similar pattern is also observed in the subset used in this study. On average, ADS utterances are longer than IDS utterances, resulting in a greater total duration of ADS speech in the final subset, as shown in Table 3.1. The selected utterances from the MB 1 corpus vary in length from approximately 1 to 10 seconds, with ADS utterances generally being longer. In terms of total duration, the ADS subset contains approximately twice as much speech as the IDS subset.

Table 3.1. *Overview of IDS and ADS clip duration statistics per speaker for the subset of the MB 1 corpus used in this study.*

Speaker	IDS Clips	IDS Total (s)	IDS Mean (s)	IDS Std (s)	ADS Clips	ADS Total (s)	ADS Mean (s)	ADS Std (s)
Speaker 1	2	3.88	1.94	0.48	2	9.44	4.72	2.44
Speaker 2	10	18.3	1.83	0.61	10	30.65	3.07	0.68
Speaker 3	10	26.08	2.61	0.84	10	86.21	8.62	4.7
Speaker 4	10	33.02	3.3	1.08	10	52.57	5.26	1.58
Speaker 5	9	30.04	3.34	1.56	9	59.61	6.62	3.94
Speaker 6	8	24.6	3.08	0.91	8	24.91	3.11	0.93
Speaker 7	8	27.4	3.42	1.66	8	39.36	4.92	2.88
Speaker 8	10	21.18	2.12	0.51	10	44.29	4.43	2.45
Speaker 9	8	22.11	2.76	1.22	8	40.67	5.08	1.98
TOTAL	75	184.5	2.75	1.18	75	347.04	5.17	3.13

3.1.2 TIMIT

A subset of TIMIT (Garofolo et al. 1993) was selected as a separate ADS dataset. The corpus consists of recordings of multiple studio-recorded read sentences from multiple US dialect regions, providing high-quality and stable adult-directed speech. The dataset provides an upper bound on reference quality for speech synthesis in this study.

Original TIMIT utterances from the same speakers were concatenated to create the reference samples for this study. As shown in Table 3.2, the selected subset contains a total of 50 concatenated reference samples. Each reference sample is from a different speaker, with an average duration of approximately 30 seconds. All references consist of speakers reading the same textual content.

Table 3.2. *Duration statistics of the subset of TIMIT dataset, Each recording is from a different speaker.*

TIMIT subset	Number of Files	Total Duration (s)	Mean Duration (s)	Std Duration (s)
All speakers	50	1539.53	30.79	2.77

3.1.3 Text Prompts for Synthesis

IDS-style text prompts were generated using GILES (Räsänen and Kocharov 2025). GILES is a computational model designed to produce age-appropriate infant-directed speech transcripts. The model used in this study was originally trained on IDS transcripts from the CHILDES database (MacWhinney and Snow 1985). GILES captures characteristic features of IDS, such as simplified syntax, shorter utterances compared to ADS, repetition of words, and age-dependent lexical choices.

In this study, GILES was used to generate ten IDS-style text prompts of approximately 50 words each, simulating language input typical for a one-year-old child. The 50-word length was chosen to provide sufficient phonetic and linguistic content while keeping synthesized outputs manageable relative to the reference speech durations. The generated prompts were manually inspected for plausibility in terms of coherence and adherence to IDS-like properties.

Adult-directed text prompts were also generated to enable comparison between synthesis based on IDS-style and ADS-style text inputs. Each prompt consisted of approximately 50 words. The ADS text prompts were generated using GPT-4 (OpenAI 2023), instructed to produce coherent adult-to-adult conversational speech while avoiding infant-directed language. Examples of both prompt types are shown in Figure 3.1.

GILES 1:

You're so daddy can wash it off. He spit out onto my computer dear.

The ring or something. Can't she bend it. I looked very pleased.

Green beans. But i tried to throw it out. Why i didn't even throw it.

QTEXT 1:

Honestly. I think the best way to move forward is by collaborating with everyone involved.

We can't just assume one approach will work. I believe communication and compromise are key.

That way we're all on the same page moving toward a common goal.

Figure 3.1. *Example of text prompts generated for synthesis using GILES and ChatGPT. The top example shows a GILES-generated prompt ("GILES 1"), and the bottom example shows a ChatGPT-generated prompt ("QTEXT 1").*

3.1.4 Preprocessing with Resemble Enhance

Reference audio from IDS datasets was preprocessed using the Resemble Enhance model (ResembleAI 2023) to improve perceptual quality prior to synthesis. Resemble Enhance consisted of two primary components: a denoiser and an enhancer.

The denoiser is a UNet-based neural network inspired by AudioSep (Liu et al. 2023). It operates on complex spectrograms and estimated both magnitude masks and phase rotations for separating speech from background noise. By applying these masks, non-speech components are suppressed while preserving the underlying speech signal.

The enhancer employs a combination of an Implicit Rank-Minimizing Autoencoder (IR-MAE) and a Conditional Flow Matching (CFM) module. The technical details of latent-

space representations and generative refinement for neural speech enhancement are introduced in the background section (Section 2.6.1). The IRMAE reconstructs speech components that may have been attenuated or lost during denoising in a learned latent space, while the CFM models probabilistic transformations between noisy and clean latent representations. The enhanced output is obtained by blending the reconstructed signal with the denoised input, in order to recover spectral detail and improve naturalness.

For the MB 1 dataset, three versions of each recording were retained: the original, a denoised-only version, and a denoised-and-enhanced version. TIMIT recordings were not processed with Resemble Enhance, as the purpose of the TIMIT dataset was to provide a high-quality baseline reference for synthesis without additional preprocessing.

3.1.5 XTTSv2 for Zero-Shot Style Transfer

XTTS-v2 (Casanova, Davis et al. 2024) is used as the TTS system for generating synthetic speech in the experiments of this study. XTTS-v2 is a zero-shot text-to-speech (TTS) model capable of synthesizing speech in the style of an unseen speaker using a short reference audio clip.

The architecture of xTTS-v2 integrates multiple components, also seen in Figure 3.2:

- **Conditioning Encoder**: extracts latent style embeddings from reference audio using a cross-attention-based architecture inspired by PerceiverIO (Jaegle et al. 2022). It maps variable-length acoustic features into a fixed set of embeddings, capturing speech characteristics such as prosody, rhythm, and timbre (Casanova, Davis et al. 2024).
- **Autoregressive Spectrogram Decoder**: based on transformer layers, predicts sequences of discrete acoustic codes from input text, conditioned on style embeddings (Betker 2023).
- **Diffusion-based Waveform Generator**: uses Denoising Diffusion Probabilistic Models (DDPMs) to refine spectrogram predictions into time-domain waveforms, improving naturalness and reducing artifacts (Ho et al. 2020).
- **Vector Quantized-Variational Autoencoder (VQ-VAE)**: discretizes mel-spectrogram frames into a finite set of latent codes for efficient representation and reconstruction of spectral patterns (Oord, Vinyals et al. 2017).
- **HiFi-GAN Vocoder**: converts spectrogram or latent acoustic representations into high-fidelity time-domain waveforms (Kong et al. 2020).
- **GPT-2-based Encoder**: models sequential structure in text and style embeddings to guide spectrogram generation (Radford, Wu et al. 2019).
- **Speaker Consistency Loss (SCL)**: encourages the model to maintain a stable

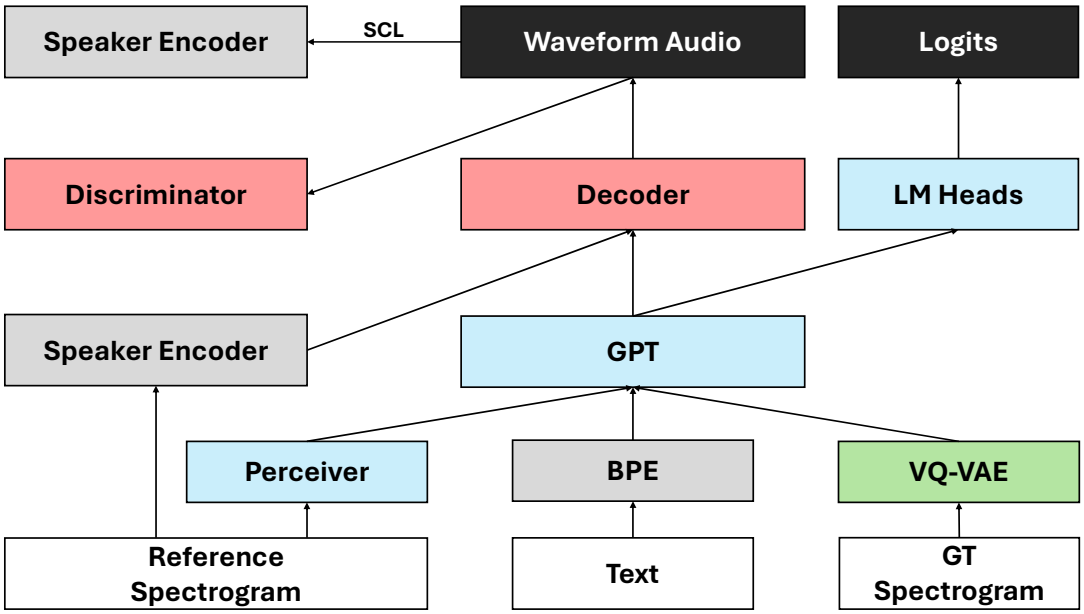

Figure 3.2. *XTTS-v2 architecture overview, reproduced from (Casanova, Davis et al. 2024). The model consists of a Vector Quantised-Variational AutoEncoder (VQ-VAE) for discrete acoustic representation, a decoder-only GPT-style transformer that predicts acoustic codes from text tokens, and a HiFi-GAN vocoder for waveform synthesis. The system is conditioned on speaker embeddings extracted from reference speech using a Perceiver-style conditioning encoder, with conditioning injected into the transformer and decoder. During training, ground-truth speech is encoded into VQ-VAE tokens, enabling auxiliary supervision through language modeling heads and a speaker consistency loss (SCL) applied on reconstructed speech.*

speaker identity across multiple synthesized utterances during training (Xin et al. 2021).

During inference, a reference audio segment is passed to the conditioning encoder to extract a latent embedding. This embedding conditions both the autoregressive spectrogram decoder and the diffusion-based waveform generator. The synthesized outputs therefore reflect both speaker identity and stylistic characteristics, such as infant-directed speech prosody.

For this study, XTTS-v2 was used to synthesize speech using IDS and ADS references from the MB 1 corpus, as well as their preprocessed versions, and ADS references from the TIMIT corpus. Each reference was paired with all IDS-style text prompts (GILES 1–10) as well as all ADS-style text prompts (QTEXT 1–10).

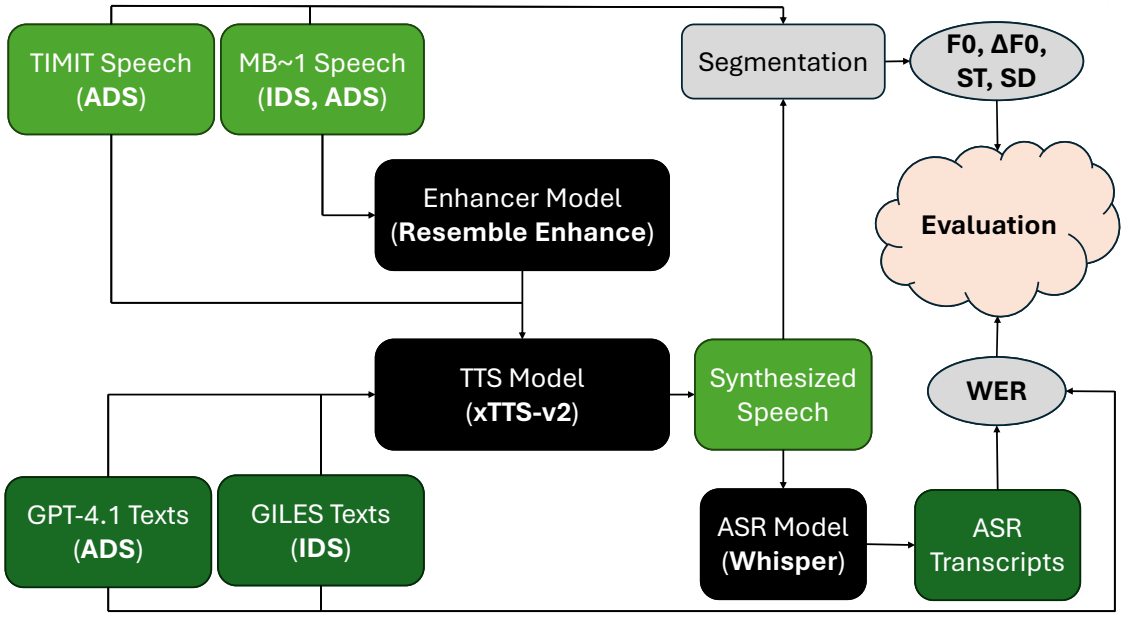

Figure 3.3. *The TTS pipeline and analysis pipeline used in the study. Subsets of the TIMIT and MB 1 datasets are used as reference speech; MB 1 is additionally processed with the enhancer model. GILES- and GPT-generated texts serve as synthesis prompts. All non-single-utterance speech samples (TIMIT references and synthesized speech) are segmented into utterances. Acoustic features (F0, delta F0, spectral tilt (ST), and syllable duration (SD)) are then measured in terms of mean and standard deviation across utterances. Word error rate (WER) is computed by comparing automatic speech recognition (ASR)-generated transcripts of synthesized speech with the original synthesis prompts.*

3.2 Analysis of Acoustic Features and Intelligibility

This section describes the analysis procedures used to quantify prosody transfer from reference to synthesized speech and intelligibility in synthesized output. The full pipeline including the analysis procedures can be seen in Figure 3.3.

The feature extraction for prosody analysis and the computation of WER are described in Subsection 3.2.2, while the segmentation procedure used to define utterance-level units is described in Subsection 3.2.1.

3.2.1 Segmentation into Utterances

All synthesized recordings and concatenated TIMIT references were segmented into utterances after synthesis. The MB 1 recordings were treated as single utterances, as they were already provided at the utterance level. The *uhm-o-meter* Praat script was

used for segmentation (De Jong et al. 2021). The script automatically detects silent and low-amplitude regions to define utterance boundaries.

Segmentation was performed to ensure that all evaluated samples corresponded to comparable utterance-level units, which enabled consistent evaluation across conditions despite differences in utterance length and pausing patterns. Segments shorter than 1 s were removed because they were predominantly caused by artifacts or non-speech components and mostly did not contain any voiced acoustic information.

3.2.2 Acoustic and Intelligibility Measures

Prosodic transfer from reference to synthesized speech was quantified using acoustic features typically differentiating in IDS and ADS, in pitch, speech rate, and spectral structure. These features were extracted at utterance level from all references and syntheses (MB 1 IDS/ADS, TIMIT, and preprocessed variants):

- **Fundamental frequency (F0)**: computed using Praat's autocorrelation-based pitch tracking. Values below 100 Hz were treated as unvoiced and excluded. A log-transformation was applied to reduce skewness and reflect perceptual scaling.
- **Delta log F0**: computed as the derivative of the estimated log F0 contour using a sliding window of five frames (step size 1), capturing local pitch dynamics.
- **Syllable duration**: estimated using ThetaSeg (Räsänen, Doyle et al. 2018), which detects syllable nuclei based on amplitude modulations. Durations (in seconds) were log-transformed.
- **Spectral tilt**: computed as the slope of a linear regression fitted to the log-magnitude spectrum of each 30 ms STFT frame (10 ms hop size).

Acoustic features were summarized at the utterance level using their mean and standard deviation. Utterance-level acoustic features were then aggregated to the recording level by computing the mean of each feature across all utterances within a recording. This resulted in a single recording-level feature vector used for subsequent analyses and visualization.

Pearson correlations (Pearson 1895) were used to quantify the correspondence between reference and synthesized speech with respect to prosody transfer. In addition, distributional differences between reference and synthesized speech due to preprocessing were analyzed using Jensen–Shannon divergence (Menéndez et al. 1997).

WER was used to assess the intelligibility of the synthesized speech. It was computed by comparing the corresponding synthesis text prompts with ASR transcriptions of the synthesized speech. Prior to calculation of WER, punctuation and special characters were removed. ASR transcriptions were generated using OpenAI's Whisper medium English-

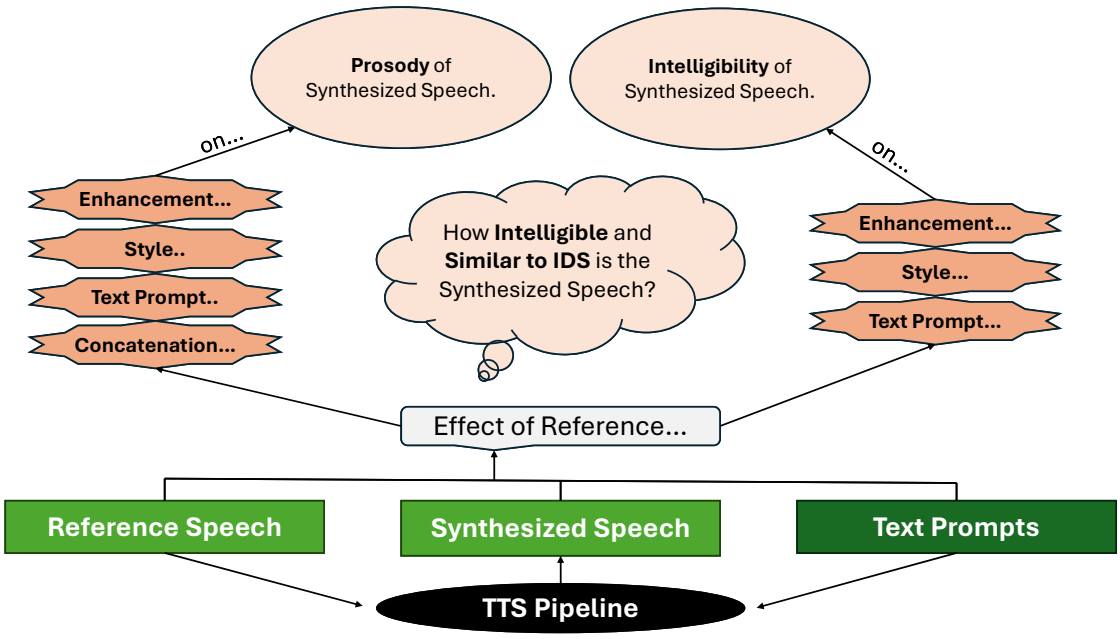

Figure 3.4. *The evaluation framework of the study. Synthesis output from the TTS pipeline is analyzed regarding prosody transfer and intelligibility. Additionally, it is examined how different modifications to reference speech (enhancement, style, concatenation) and text prompts affect the resulting synthesis in terms of intelligibility and prosody transfer.*

only model (Radford, Kim et al. 2022), which performs speech-to-text conversion using an encoder–decoder transformer architecture.

3.3 Experimental Design

This section describes the experimental design used to analyze how systematically manipulated input factors influence prosody transfer and intelligibility in xTTS-v2-based speech synthesis.

The synthesis procedure and input construction are defined in Section 3.1, which specifies how reference recordings and text prompts are combined to generate synthetic speech. Building on this procedure, the present study organizes the resulting outputs within a controlled multi-factor experimental design. This evaluation framework is presented in Figure 3.4.

The experiment consists of two main components: the evaluation of prosody transfer from reference to synthesis and the evaluation of the intelligibility of the synthesized speech.

The prosody transfer evaluation is structured around four main factors:

1. Reference speech style (IDS vs. ADS)
2. Preprocessing condition (original, denoised, enhanced)
3. Text prompt style (IDS vs. ADS)
4. Reference construction method (natural vs. concatenated utterances)

The MB 1 corpus serves as the primary dataset and is used for both prosody transfer and intelligibility evaluations, while the TIMIT corpus is included as an external benchmark dataset used exclusively for intelligibility evaluation. The synthesized speech from the MB 1 corpus is analyzed using acoustic feature statistics to evaluate prosody transfer, whereas intelligibility is assessed using Word Error Rate (WER) for synthesized speech from both the MB 1 and TIMIT corpora.

3.3.1 Effect of Style on Prosody Transfer from Reference to Synthesis

The evaluation factor focuses on whether a zero-shot TTS model transfers IDS-specific prosodic and acoustic characteristics from natural reference recordings into synthesized speech, and on how reference speaking style influences this transfer within the defined experimental design.

The analysis considers two complementary comparisons under otherwise identical synthesis conditions:

1. **Reference-to-synthesis comparison:** reference recordings are compared with their corresponding synthesized outputs for IDS and ADS, while holding text prompts and speaker identity constant.
2. **Cross-style synthesis comparison:** synthesized IDS and ADS outputs are compared while holding speaker identity and text prompt constant.

Together, these comparisons isolate the effect of reference speech style on prosody transfer while controlling for other experimental factors.

3.3.2 Effect of Preprocessing on Prosody Transfer from Reference to Synthesis

This evaluation factor examines how preprocessing of reference speech affects prosody transfer in xTTS-v2-based synthesis under otherwise identical conditions.

The analysis considers three controlled comparisons:

1. **Reference-level effects:** comparison of original and preprocessed reference recordings to quantify direct acoustic changes induced by preprocessing.

2. **Synthesis-level effects:** comparison of synthesized speech generated from original and preprocessed references to evaluate how reference enhancement influences the acoustic properties of the generated speech relative to the original reference signal.
3. **Reference–synthesis distributions:** comparison of reference and synthesis distributions under original and preprocessed conditions to assess whether relative prosodic structure is preserved after preprocessing.

3.3.3 Effect of Text Prompt Style on Prosody Transfer from Reference to Synthesis

This evaluation factor examines whether linguistic style in the text prompt influences prosody transfer in zero-shot TTS under otherwise identical synthesis conditions.

Two prompt conditions are compared:

- IDS-style prompts (GILES)
- ADS-style prompts (QTEXT)

All other synthesis parameters, including reference speech, speaker identity, and dataset, are held constant, allowing for isolation of text-driven effects.

3.3.4 Effect of Reference Concatenation on Prosody Transfer from Reference to Synthesis

This evaluation factor examines the effect of reference speech construction on synthesis, motivated by the xTTS-v2 requirement that reference audio clips should exceed 6 seconds in duration, defined in (Coqui 2024), whereas the reference utterances used in this study are often shorter (Section 3.1).

Two conditions are defined to satisfy this constraint: natural (non-concatenated) utterances and concatenated utterances constructed from shorter IDS segments in the MB 1 dataset. This setup enables a direct comparison between natural and artificially extended reference speech while controlling for speaker identity and text prompt conditions.

Synthesized outputs generated from these two reference construction methods are compared to assess their impact on prosody transfer in xTTS-v2.

3.3.5 Intelligibility of the Synthesized Speech

This evaluation assesses the intelligibility of synthesized speech, operationalized using word error rate (WER) computed from automatic speech recognition outputs.

The analysis includes two datasets: MB 1 as the primary dataset used for all synthesis conditions, and TIMIT as an external benchmark used exclusively for robustness evaluation of intelligibility findings.

Comparison is conducted across three synthesis types (enhanced MB 1 IDS, enhanced MB 1 ADS, and TIMIT). For each synthesis type, both IDS-style (GILES) and ADS-style (QTEXT) text prompts are applied, resulting in a fully crossed design with six conditions:

- MB 1 IDS synthesis with IDS prompts
- MB 1 IDS synthesis with ADS prompts
- MB 1 ADS synthesis with IDS prompts
- MB 1 ADS synthesis with ADS prompts
- TIMIT synthesis with IDS prompts
- TIMIT synthesis with ADS prompts

Word error rates are compared across conditions to evaluate intelligibility differences under varying synthesis configurations.

Additionally, synthesized speech generated from original and enhanced references using IDS-style prompts is compared to examine the effect of reference enhancement on intelligibility. The corresponding WER results are contextualized with respect to Whisper's performance across different speech corpora.

4. RESULTS

This chapter reports the experimental results on prosody transfer and intelligibility of the synthesized speech. First, Section 4.1 analyzes the effects of speech style, text prompt style, and reference preprocessing on prosody transfer, followed by an analysis of the effect of reference concatenation in Subsection 4.1.4. Second, the intelligibility of the synthesized speech is evaluated in Section 4.2.

4.1 Transfer of Prosody from Reference to Synthesis

First, the effect of reference style on prosody transfer is examined (Subsection 4.1.1), followed by the effect of preprocessing (Subsection 4.1.2), the effect of text prompt style (Subsection 4.1.3), and finally the effect of reference concatenation (Subsection 4.1.4).

4.1.1 Effect of Reference Style

Figure 4.1 shows correlations between original reference speech and the corresponding synthesized outputs. Overall, correlation patterns are similar for IDS and ADS styles. The strongest correlations are observed for mean F0 and spectral tilt in both speaking styles. In contrast, syllable duration (mean and standard deviation) and the standard deviation of delta F0 consistently show weak correlations. Standard deviation-based measures of these features are not statistically significant for either IDS or ADS.

Most other acoustic features exhibit moderate correlations, with relatively small differences between IDS and ADS. The largest style-related difference is observed for mean delta F0, where ADS shows higher correlation values ($r = 0.25$) than IDS ($r = 0.08$). Overall, standard deviation-based features tend to show weaker correlations than mean-based features across both speaking styles.

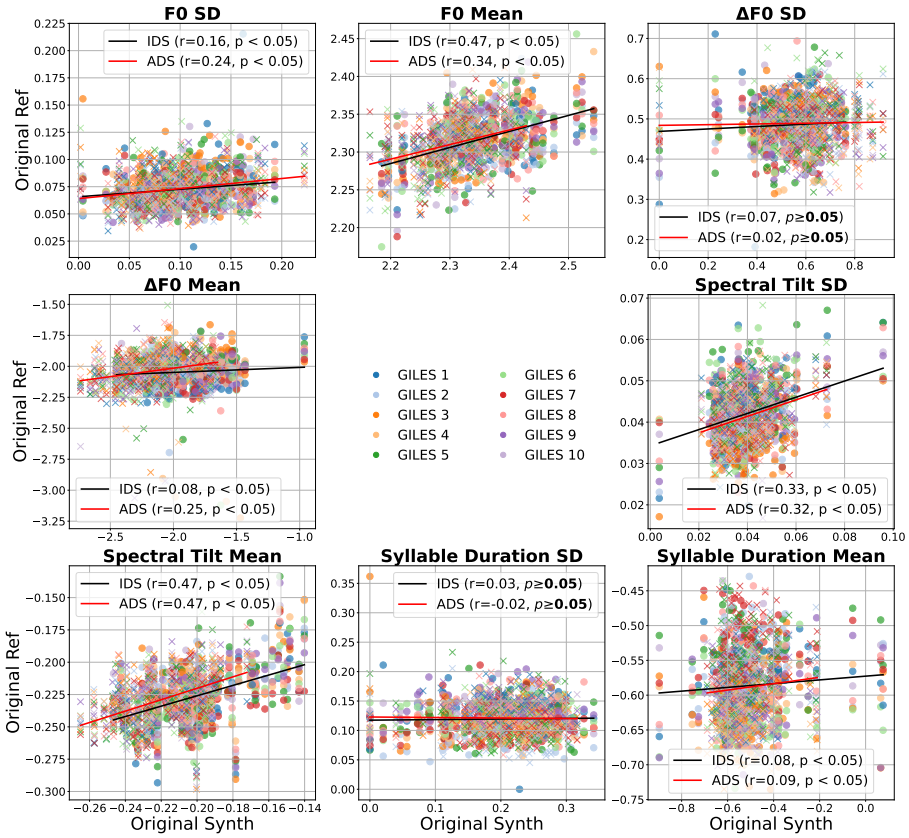

Figure 4.1. *Scatter plots with linear regression fits comparing IDS (red) and ADS (blue) original reference speech to their corresponding synthesized outputs in terms of mean and standard deviation across four acoustic metrics (log F0, Δ log F0, spectral tilt, and log syllable duration). Different IDS text prompts (GILES) are indexed and color-coded (GILES 1–10) to highlight potential prompt-dependent variation.*

Figure 4.2 shows IDS and ADS synthesized speech plotted together with the corresponding speakers and text prompts. Overall, correlation values are higher for reference–synthesis pairs (Figure 4.1) than for IDS–ADS synthesis comparisons (Figure 4.2) across most acoustic features.

Exceptions are observed for mean and standard deviation of syllable duration, as well as the standard deviations of spectral tilt and delta F0, where IDS–ADS synthesis correlations are comparable or slightly higher. Delta F0 variability remains low in both comparison types but is statistically significant in the IDS–ADS synthesis comparison.

Mean delta F0 and F0 standard deviation are not statistically significant in IDS–ADS synthesis comparisons, whereas both are significant in reference–synthesis pairs. For mean F0, correlations are again stronger in reference–synthesis pairs than in IDS–ADS synthesis comparisons.

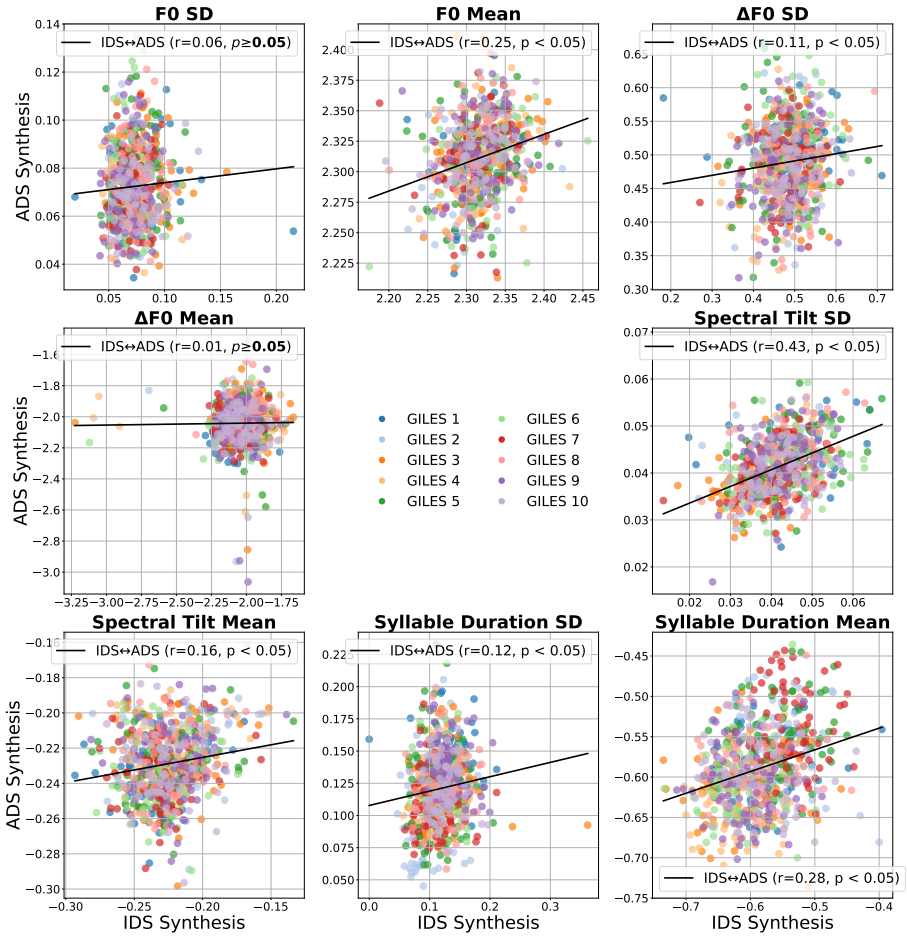

Figure 4.2. *Scatter plot with linear regression fit (black) of paired IDS—ADS syntheses matched by speaker and GILES text index (1–10). Each synthesized sample is included once in the comparison.*

4.1.2 Effect of Reference Preprocessing

Figure 4.3 and Table 4.1 compare kernel density estimates (KDEs) of acoustic feature distributions in synthesized speech generated from differently preprocessed references against those of the original reference speech, where the KDEs are estimated across all the corresponding of each category. The results show that both denoising and enhancement have only a marginal effect on distributional similarity. The enhanced condition yields the lowest Jensen–Shannon divergence; however, the difference compared to the original condition is minimal, indicating a negligible effect. The denoised condition shows slightly higher divergence than the unprocessed condition.

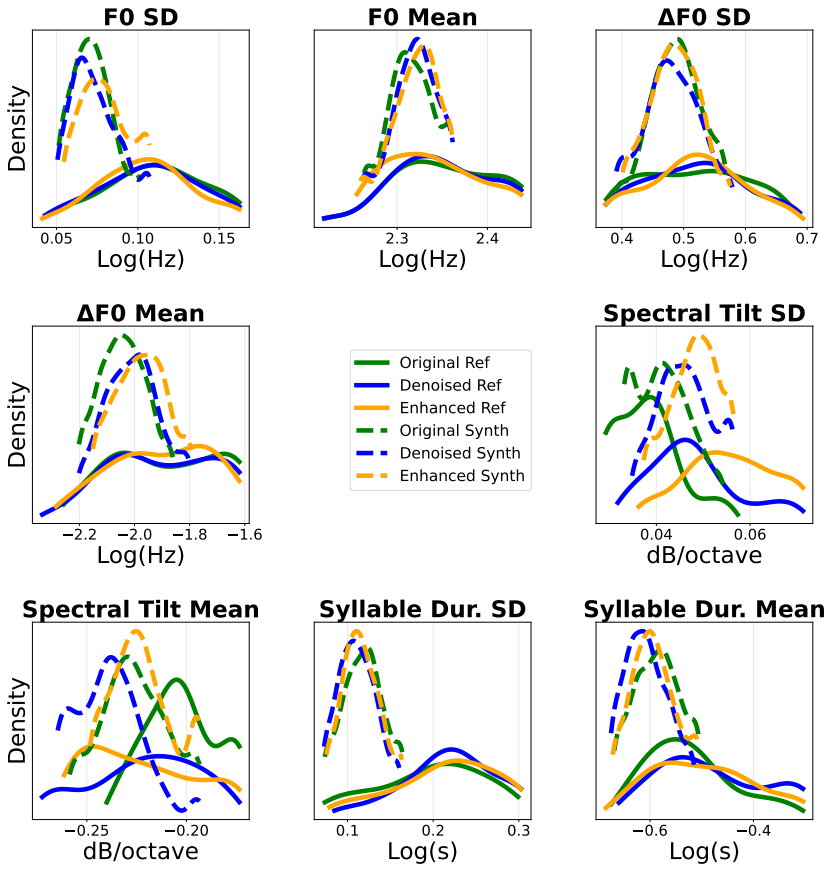

Figure 4.3. *Kernel density estimates (KDEs) of acoustic features comparing original reference speech and synthesis produced from references with different preprocessing conditions. Colors and line types distinguish preprocessing techniques (green for original, blue for denoised, and yellow for enhanced). Dashed lines represent synthesis, while solid lines represent original references. Values are filtered to the 5th–95th percentile range to reduce the influence of outliers.*

Table 4.1. *Jensen–Shannon distances between kernel density estimates (KDEs) of synthesized speech and original reference utterances. Synthesized speech is generated from references processed under three conditions: no processing (original), denoised, and denoised + enhanced. The table reports mean and standard deviation values aggregated across all acoustic features, providing an overall measure of divergence from the original reference distributions. Lower values indicate higher similarity to the original distributions, while higher values indicate greater divergence.*

Aggregated Distances to Original References		
Feature aggregation	**Preprocessing type**	**Jensen–Shannon**
All features (mean of means)	Denoised Synthesis	0.2936
All features (mean of means)	Enhanced Synthesis	0.2646
All features (mean of means)	Original Synthesis	0.2686
All features (mean of standard deviations)	Denoised Synthesis	0.3239
All features (mean of standard deviations)	Enhanced Synthesis	0.3391
All features (mean of standard deviations)	Original Synthesis	0.3152

It can be seen from Table 4.1 that Jensen–Shannon divergence values remain within a narrow range (approximately 0.25–0.35), indicating moderate distributional differences between synthesized and reference speech, regardless of preprocessing condition. Figure 4.3 shows that, while feature distributions largely overlap across conditions—particularly with respect to the x-axis location of the mean acoustic features—there are consistent observable differences. Most differences arise from synthesis distributions being more concentrated and less dispersed in general, rather than from variation induced by reference preprocessing. Spectral tilt exhibits the highest similarity across conditions.

Figure 4.4 shows correlations between original and enhanced reference speech. High correlation values are observed across all acoustic features for both IDS and ADS, indicating that neural enhancement does not substantially alter the measured acoustic properties of the references. This pattern holds consistently across both speaking styles.

Figure 4.5 shows correlations between original reference speech and synthesized outputs generated from enhanced references. Compared with Figure 4.1, correlation patterns remain largely similar across synthesis conditions. In some features, enhanced synthesis yields slightly higher correlations, while in others correlations decrease.

For IDS and ADS, mean F0 shows a slight decrease in correlation under enhancement, while its standard deviation shows a small increase. A similar but inverse pattern is observed for spectral tilt, where mean values increase, but standard deviation weakens and becomes non-significant for IDS under enhanced synthesis. In contrast, the standard

deviation of delta F0 becomes statistically significant under enhanced synthesis but not under original synthesis. Syllable duration, both mean and standard deviation, remains weakly correlated and non-significant in both synthesis conditions.

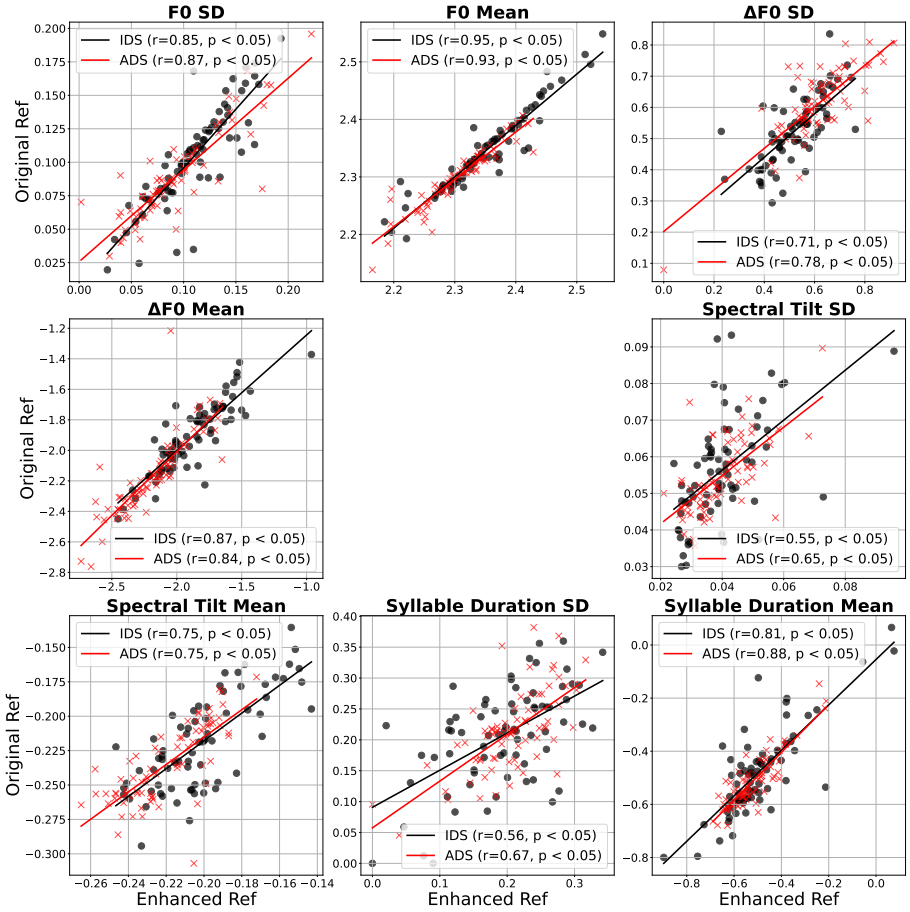

Figure 4.4. *Scatter plots with linear regression fits comparing original reference speech to corresponding enhanced reference speech for IDS (black) and ADS (red).*

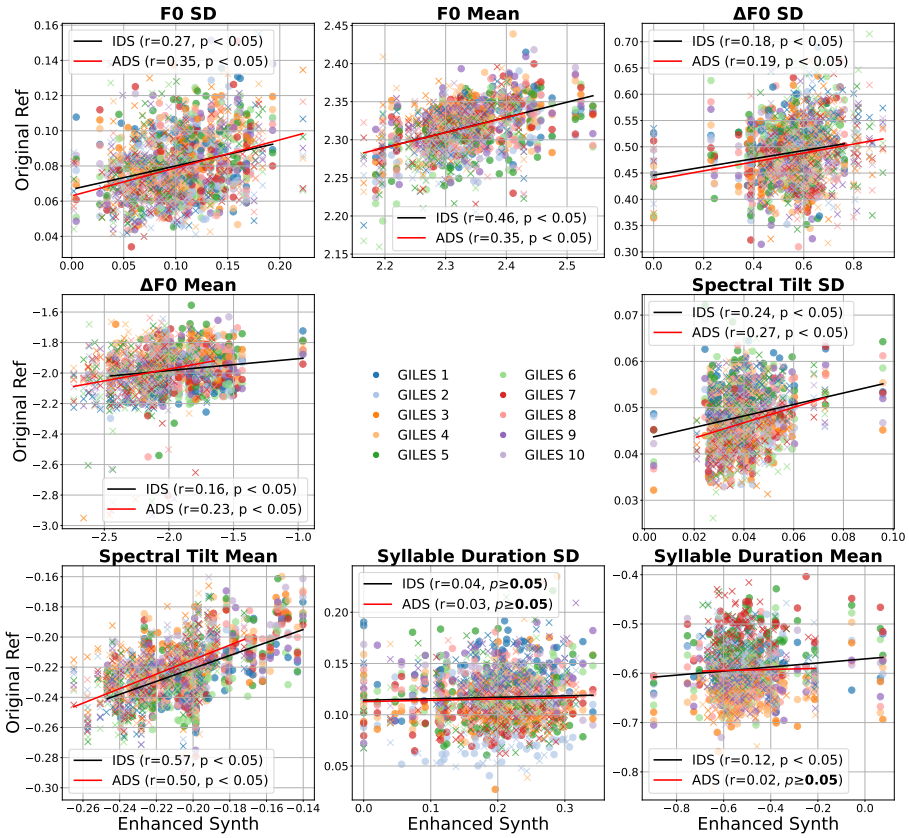

Figure 4.5. *Scatter plots with linear regression fits comparing IDS (red) and ADS (blue) original reference speech to synthesized outputs generated from the corresponding enhanced references. IDS text prompts (GILES) are indexed and color-coded (GILES 1–10) to highlight potential prompt-dependent variation.*

4.1.3 Effect of Text Prompt Style

Figure 4.6 shows synthesis generated using ADS-style text prompts (QTEXT). Compared to Figure 4.1, which uses IDS-style text prompts (GILES), both QTEXT- and GILES-based synthesis exhibit similar correlation patterns with the reference speech.

Table 4.2, which reports aggregated r-values for both prompt styles, further confirms that correlations are closely matched across conditions. QTEXT-based synthesis yields slightly stronger correlations for both IDS and ADS.

Table 4.2. *Aggregated statistically significant correlations ($p < 0.05$) between original reference recordings and synthesized outputs. Synthesis is driven by two text prompt sources: IDS-style prompts generated by GILES and ADS-style prompts generated by ChatGPT (QTEXT).*

Style	Aggregation	r GILES	r QTEXT
IDS	All features (mean of means)	0.328	0.372
IDS	All features (mean of standard deviations)	0.188	0.223
ADS	All features (mean of means)	0.347	0.421
ADS	All features (mean of standard deviations)	0.181	0.205

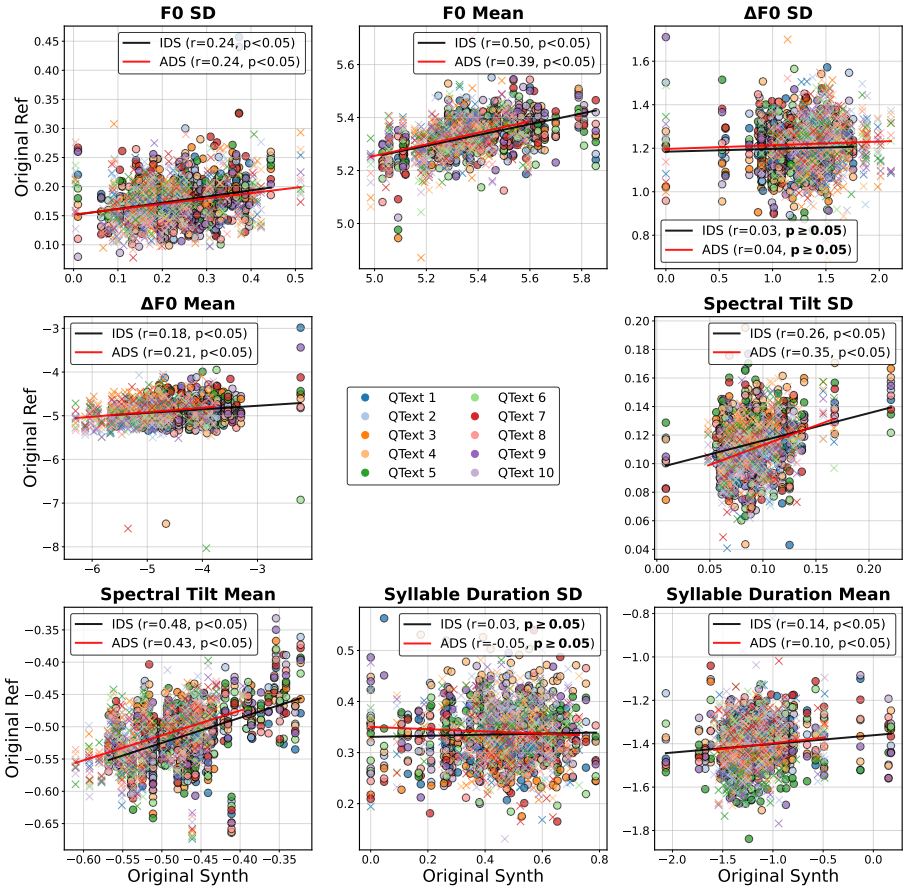

Figure 4.6. *Scatter plot with linear regression fits comparing IDS (blue) and ADS (red) reference speech with their corresponding non-enhanced synthesized outputs using ADS-style text prompts (QTEXT). Text prompts are indexed as QTEXT 1–10 to highlight potential variation due to prompt differences.*

4.1.4 Effect of Reference Concatenation

Figure 4.7 compares original IDS references (utterance-based synthesis) and concatenated versions of the same references, together with the corresponding synthesized out-

puts (concatenated synthesis). The results show that utterance-based synthesis consistently yields stronger correlations with synthesized speech across features compared to concatenated synthesis.

In addition, F0 standard deviation and mean delta F0 are not statistically significant under the concatenated condition. Syllable duration and delta F0 exhibit weak correlations overall, with the corresponding standard deviations also not statistically significant regardless of concatenation.

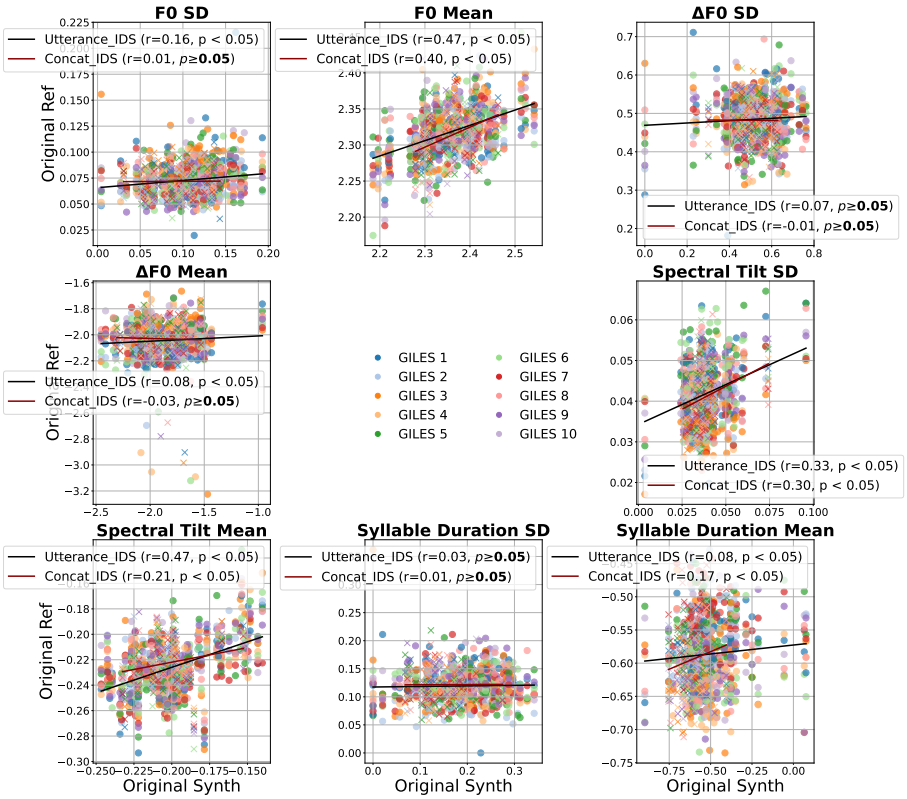

Figure 4.7. *Scatter plots with linear regression fits comparing IDS-synthesized speech generated from individual utterances versus concatenated inputs. Text prompts are labeled GILES 1–10 to highlight potential variation introduced by differences in the prompts.*

4.2 Intelligibility of the synthesized speech

Figure 4.8 shows a comparison of word error rate (WER) results for enhanced and original ManyBabies 1 (MB 1) references using IDS-style text prompts. Synthesis with enhanced references produces more intelligible speech (WER $\approx$ 11%) compared to non-enhanced references (WER $\approx$ 19%).

Figure 4.9 shows WER distributions across IDS and ADS text prompts and reference conditions. A clear effect of text prompt style is observed: speech synthesized from ADS-style text prompts (QTEXT) yields substantially lower WER than speech synthesized from IDS-style text prompts (GILES). WER for GILES-based synthesis is approximately an

order of magnitude higher than for QTEXT-based synthesis, whereas ADS-style prompt synthesis produces very low error rates (WER $\approx$ 1

When considering specific dataset combinations, GILES prompts paired with IDS-style references yield slightly lower WER than when paired with TIMIT ADS audio references. For MB 1 ADS data with GILES prompts, WER differences are marginal compared to IDS audio with GILES prompts.

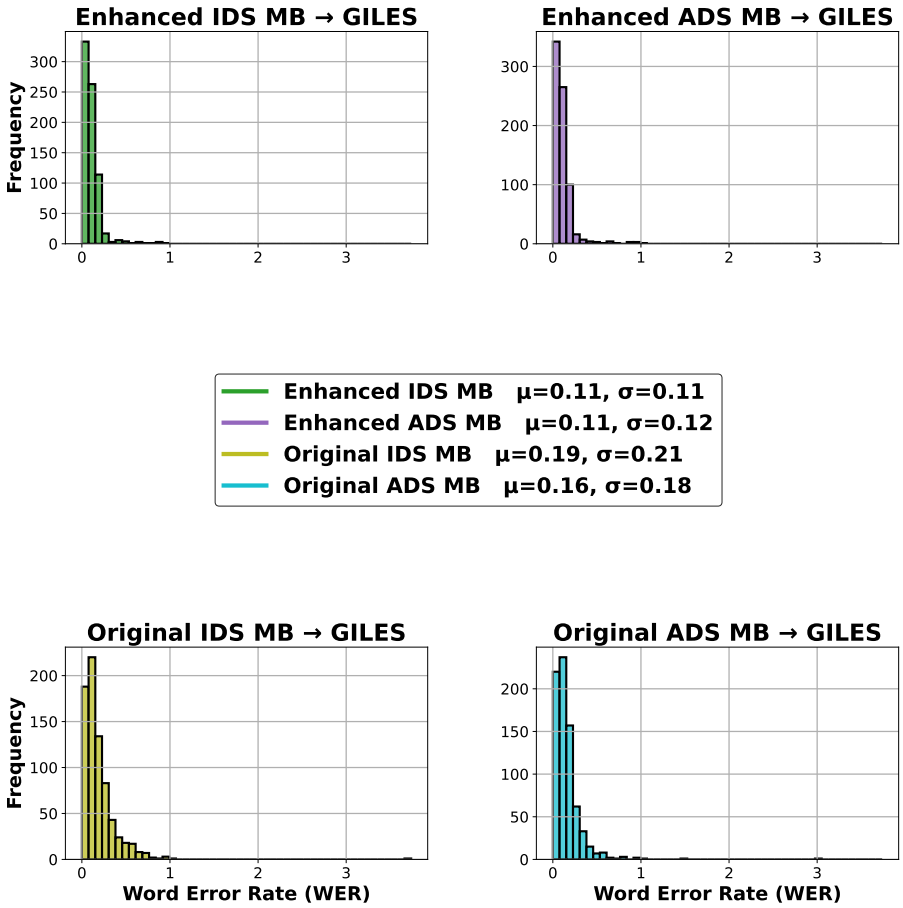

Figure 4.8. *Histograms comparing WER distributions for synthesis generated from original and enhanced MB 1 references using GILES text prompts. Each subplot reports the mean (μ) and standard deviation (σ) of WER.*

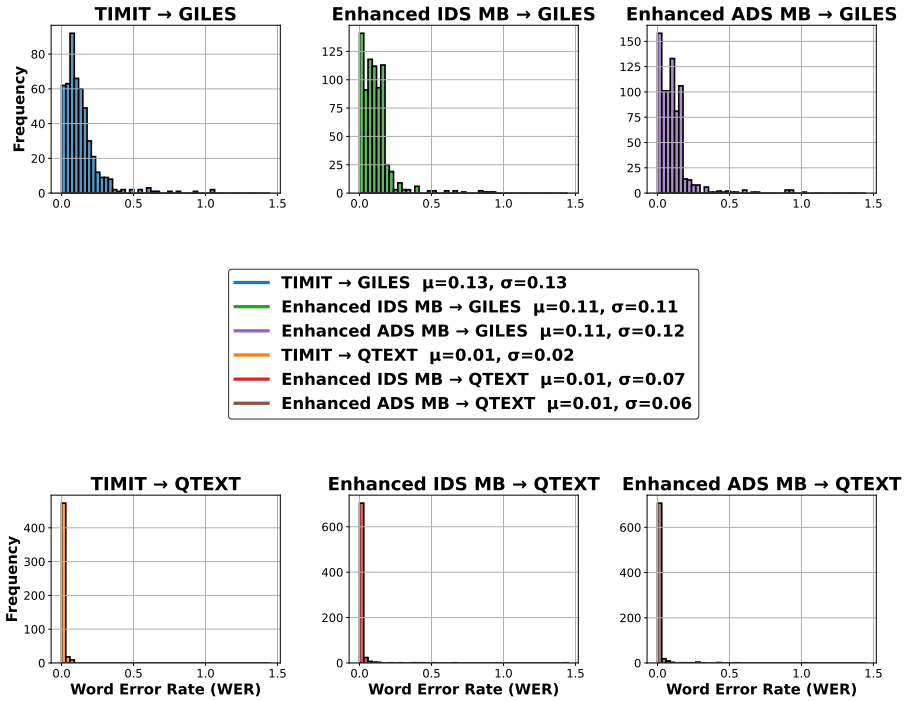

Figure 4.9. *Histograms showing the distribution of WER across different synthesis–text type combinations. Each subplot corresponds to a specific pairing of dataset (TIMIT, MB 1) and text prompt type (GILES, QTEXT). WER values are derived from Whisper transcription outputs and aggregated per condition. Each panel reports the mean (μ) and standard deviation (σ) of WER.*

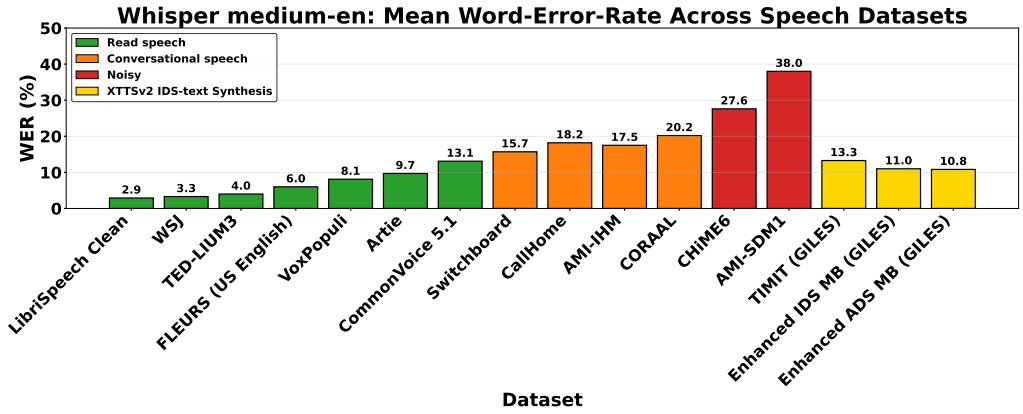

Figure 4.10. *Bar chart showing mean WER across speech datasets (read, conversational, and noisy speech) alongside xTTS-v2 IDS-text synthesis conditions (TIMIT, IDS MB, ADS MB). Adapted from Whisper results for different datasets documented in Radford et al. (Radford, Kim et al. 2022).*

Overall, WER values are low, indicating that the synthesized speech generated with xTTS-v2 is generally intelligible. Compared with reported Whisper performance across datasets in Figure 4.10, synthesized speech from enhanced IDS references achieves WER levels comparable to semi-controlled datasets such as Artie (Meyer et al. 2020) and VoxPopuli (Wang, Riviere et al. 2021). Speech synthesized using IDS-style references

and IDS-style text prompts yields slightly higher error rates than these datasets, but lower error rates than Common Voice (Ardila et al. 2020).

5. DISCUSSION

The aim of this work was to enable natural-sounding, realistic IDS synthesis using a zero-shot TTS system. The synthesized speech generated by the proposed approach was evaluated from several perspectives.

The results are first discussed in terms of the transfer of prosodic characteristics from reference speech to synthesis (Section 5.1), followed by the intelligibility results (Section 5.2). Finally, the study's limitations (Section 5.3), ethical considerations (Section 5.4), and directions for future work (Section 5.5) are outlined.

5.1 Transfer of IDS-related acoustic characteristics

The results indicate that infant-directed speech (IDS) and adult-directed speech (ADS) characteristics are incorporated in a comparable manner during synthesis, without evidence of systematic bias toward either style. This results in distinct but similarly represented synthetic IDS and ADS outputs. However, overall prosody transfer from reference to synthesis varies across features, with some showing stronger correspondence than others.

Mean fundamental frequency shows strong preservation between reference audio and synthesized samples across multiple analyses. This is particularly relevant for IDS synthesis, as prior work has identified pitch as one of the most prominent characteristics of IDS (Cox et al. 2023). The standard deviation of F0 also transfers to some extent, although less consistently than the mean.

In contrast, delta F0 does not show clear transfer from reference to synthesis, regardless of speaking style. It also does not exhibit correspondence between IDS and ADS synthesis conditions, despite the two differing only in reference style.

Delta F0 is intended to capture short-term changes in pitch direction and thus reflects dynamic intonation patterns often associated with IDS compared to ADS. However, it is also highly sensitive to linguistic content. Since reference and synthesized datasets differ in both content and duration—where reference utterances are often very short (e.g., IDS segments under 3 seconds in the ManyBabies 1 (MB 1) dataset), while synthesized utterances can exceed 30 seconds—the observed Delta F0 distributions are likely influenced

more by differences in speech material than by stylistic factors alone. This suggests that Delta F0 may not be robustly estimated in the current experimental setup.

Spectral tilt shows evidence of transfer from reference to synthesized speech and also exhibits correspondence between IDS and ADS synthesis. This is expected, as spectral tilt is a strongly speaker-dependent acoustic feature and speaker identity is held constant across conditions. However, mean spectral tilt shows stronger correlation in reference–synthesis comparisons than in IDS–ADS comparisons, suggesting that style-related properties of the reference speech still influence the synthesized output, but not in a strictly equivalent manner across styles.

Syllable duration shows consistently weak correspondence between reference and synthesized speech, indicating that the model does not reliably transfer syllable-level temporal structure. This pattern holds for both IDS and ADS synthesis and does not appear to depend on either reference style or text prompt condition.

Overall, these findings suggest that syllable duration is not consistently conditioned on the reference signal in the current synthesis pipeline. This may reflect limitations in how fine-grained temporal structure is encoded during generation. Given that both IDS and ADS synthesis show limited correspondence, the model may not reliably encode syllable-level duration information from the reference audio. Further evaluation with alternative zero-shot TTS models would help determine whether this limitation is model-specific or more general.

Across features, standard deviations tend to correlate more weakly than mean-based measures, suggesting that the model produces smoother and more uniform acoustic realizations than those observed in natural reference speech. Finally, enhanced references show strong correspondence with original recordings, indicating that enhancement preserves IDS-specific characteristics of spectral tilt, particularly in both mean and variability.

5.1.1 Effect of reference audio enhancement on the transfer of prosody

The results indicate that neural enhancement does not strongly affect the acoustic feature distributions of the reference recordings. F0-related features are only minimally affected, as reflected in correlations between original and enhanced references. The most notable differences are observed in spectral tilt, which is expected, since denoising modifies the spectral frequency distribution by attenuating higher-frequency components.

With respect to synthesized speech, reference enhancement leads to improved correspondence between reference and synthesis. This is reflected in stronger correlations for F0 standard deviation, delta F0 statistics, and mean spectral tilt. These findings suggest that enhanced references yield synthesized speech that more closely reflects IDS-style

prosodic characteristics, while maintaining overall acoustic feature distributions similar to those of the original IDS references.

In contrast, denoising alone produces greater divergence between the acoustic feature distributions of synthesized speech and the original references compared to the unprocessed condition. One possible explanation is that neural denoising alters the spectral structure of the reference signal, while the synthesis model does not explicitly reconstruct such modified spectral characteristics and instead generates speech with a more typical spectral profile. This mismatch may increase differences between reference and synthesized speech, particularly in spectral tilt.

It should be noted that the MB 1 references are relatively high-quality compared to typical naturalistic IDS datasets. As discussed in Subsection 2.2.2, such datasets are often collected in natural infant-caregiver environments, which lack the controlled recording conditions of the MB 1 dataset (Frank et al. 2020). Consequently, preprocessing may yield even greater benefits when applied to lower-quality datasets.

5.1.2 Effect of text prompt style and reference audio concatenation on prosody transfer

The text prompts do not appear to produce a measurable difference in prosody transfer from reference to synthesized speech.

Reference concatenation, however, results in generally weaker correlations and introduces statistical non-significance for several features. It should be noted that concatenated references have a smaller effective number of independent data points for estimating correlations between datasets, which may increase sensitivity to outliers in the statistical evaluation. Although concatenation satisfies the reference duration requirement of XTTSv2 (Casanova, Davis et al. 2024), it does not improve prosody transfer performance.

5.2 Intelligibility of the Synthesis

Results from the ASR word error rate (WER) analysis suggest that speech intelligibility is more strongly influenced by linguistic content than by speaking style in the reference audio. However, enhancing the reference audio leads to lower WER in the synthesized speech compared to the original references, indicating improved intelligibility. Additionally, syntheses generated from IDS-style text prompts exhibit higher WER than those generated from ADS-style prompts. IDS-style text prompt synthesis shows intelligibility comparable to established read-speech datasets, though below high-fidelity or studio-quality levels when evaluated using Whisper (Radford, Kim et al. 2022).

Synthesis from enhanced IDS references yields slightly lower WER than synthesis from

ADS references from TIMIT when using identical IDS-style prompts. With ADS-style prompts, the results are comparable in terms of WER. Synthesis using ADS references from the MB 1 dataset follows a similar pattern, but only to a very small extent. While the observed differences are small, they suggest that IDS references do not degrade intelligibility and may potentially support more accurate synthesis of child- or infant-directed language content. However, it should be noted that the TIMIT references were concatenated. As discussed in Subsection 5.1.2, concatenated references do not perform as well with as natural utterances. This concatenation may negatively affect the intelligibility of the generated speech, similarly to its observed impact on prosody transfer.

Qualitative listening indicates that words more typical of IDS, such as “peekaboo,” “yucky,” “sweetie,” or “doggie,” may be articulated more clearly and with prosodic patterns closer to natural IDS when synthesized from IDS references compared to ADS references within the same speech samples. This effect may reflect the model’s learned mapping between textual patterns and IDS-like prosody during training. However, no formal perceptual evaluation was conducted to quantify this effect, and therefore these observations remain speculative.

The higher WER observed for IDS-style text prompts compared to ADS prompts may reflect reduced robustness of the ASR system to infant-directed linguistic content, due to its atypical structure and vocabulary relative to the data on which the system was trained. Notably, ADS prompts produced virtually no errors, which is lower than even high-quality reference datasets evaluated with Whisper (approximately 1% error compared to 2–9% for LibriSpeech). However, based on subjective listening, the TTS model does not appear to have difficulties pronouncing words such as “peekaboo” or “yucky,” which may be less frequent in conventional adult-directed speech. The IDS-specific words and linguistic structure also appear subjectively intelligible. The observed errors may therefore be more closely related to the ASR system itself, in particular Whisper, and its robustness to IDS-style linguistic input.

It should be noted that a more systematic evaluation would be required to isolate the source of transcription errors. Future work should include ASR evaluation of original IDS recordings, comparison across multiple ASR systems, and human perceptual studies.

5.3 Study Limitations

Several limitations of the current study should be acknowledged. First, the comparisons between IDS and ADS are limited by the sample size available per speaker in the MB 1 dataset. In addition, the acoustic metrics measured in this study inherently reflect both speaker-specific characteristics and stylistic properties, making it difficult to disentangle the two when comparing different speakers without a substantial amount of speech data per speaker.

More specifically, while care is taken to match the number of utterances per speaker in the reference dataset, it does not contain balanced amounts of speech across styles. The IDS condition contains approximately half the amount of data (in terms of duration) compared to ADS. A more reliable analysis would require a larger number of utterances per speaker or a more sophisticated statistical model that accounts for the amount of data per speaker as a covariate. This could potentially yield more robust estimates of how speaking style affects prosody transfer.

This study focuses exclusively on English. Although acoustic characteristics of IDS have been shown to share broad cross-linguistic similarities (Cox et al. 2023), the performance of the xTTS-v2 model may not directly transfer to other languages despite its multilingual design. More generally, due to the evaluation being limited to a single IDS corpus, the generalizability of the findings remains uncertain.

In particular, it is unclear to what extent the observed results from the experiment are specific to xTTS-v2, given the limited transparency regarding its training data. Evaluating additional models with different training corpora and architectural designs would be necessary to determine whether the observed results are model-specific or reflect more general properties of modern TTS systems.

5.4 Ethical considerations

In an ideal case, TTS development for IDS would use audio datasets collected from naturalistic environments involving young children. However, almost none of the existing datasets were originally collected with generative voice modeling in mind—either because such models did not exist at the time of data collection or because this type of technological use was outside the scope of the original studies. Due to the lack of informed consent from participants for such secondary use, it would be ethically questionable to employ these data for training neural TTS models. The same considerations apply to their use as reference speech in a zero-shot TTS setting. Even if current zero-shot TTS models are not capable of directly reproducing speaker identity from reference audio and primarily transfer more general approximations of the entangled speaker–style properties of the reference, future models might be capable of direct identity transfer. Identity transfer is, after all, part of what models such as xTTS-v2 are designed to achieve, namely "transferring the voice characteristics of a speaker" (Casanova, Davis et al. 2024).

Although the current approach of this study avoids training on these data and focuses solely on investigating style transfer rather than speaker identity, using publicly available MB 1 data for which TTS use permissions were obtained from the original data collectors, the system is not guaranteed to prevent identity-related characteristics from being reproduced in synthetic speech.

Hence, whether performing model training, using reference-encoder-based TTS, or sharing data produced by such systems, publicly available resources and tools should rely on IDS data that has been specifically collected for this purpose (or with otherwise broad use consents, as in the case of MB 1), or alternatively on ensuring post-hoc informed consent from the participants.

Human voice recordings may constitute personal data and, when processed in a manner that enables the identification of an individual, fall under the category of biometric data as defined in the General Data Protection Regulation (GDPR) (*Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation), Article 4(14)* 2016) and within biometric recognition frameworks (see Jain et al. 2004 for a review). Biometric data are considered a special category of personal data, the processing of which is generally prohibited unless specific legal conditions are fulfilled (*Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation), Article 9(1)* 2016). In the context of modern neural speech synthesis, even relatively small speech samples may enable the generation of large amounts of synthetic speech that preserve speaker-related characteristics, increasing the potential risks associated with the misuse or redistribution of such data. Consequently, processing speech recordings without an appropriate legal basis, including explicitly informed consent where required, may constitute a violation of applicable data protection legislation. For this reason, the collection, use, and distribution of speech data in neural speech processing and generative speech synthesis research must be approached with particular care. From both a legal and academic perspective, the irresponsible use of speech data without appropriate permissions would violate established legal frameworks and fall short of accepted standards of responsible research practice.

5.5 Future work

Future research could address the limitations of the study in several ways. Automated selection of IDS reference samples based on metrics such as signal-to-noise ratio (SNR) and the prominence of IDS-related pitch tendencies could better quantify the degree of prosody transfer from reference to synthesis by providing more representative references of IDS and ADS styles, and reduce subjective bias in selecting representative reference audio samples. Additionally, the issue of syllable duration not transferring reliably should be thoroughly investigated, possibly by testing other TTS models than xTTS-v2. This would allow determining whether the problem is common to reference-encoder-based models or specific to the xTTS-v2 architecture.

Formal perceptual evaluations involving human listeners would also provide valuable complementary insight into how synthesized IDS is perceived relative to natural IDS, and how synthesized IDS is perceived compared to ADS. Finally, extending the analysis to

additional datasets and TTS architectures would help determine the generalizability of these findings.

Further testing should be conducted with a dataset gathered in a more natural environment, if such data were to become available. This would also allow testing how well the results of this study generalize to other datasets.

6. CONCLUSION

This study investigated whether a reference-conditioned zero-shot text-to-speech (TTS) system, xTTS-v2 (Casanova, Davis et al. 2024), can synthesize intelligible infant-directed speech (IDS) when conditioned with IDS reference audio and text inputs consisting of IDS-style text prompts. The results demonstrate that, when conditioned on IDS references, the model preserves key IDS-related acoustic characteristics, particularly mean pitch and spectral tilt, while maintaining speech intelligibility comparable to that achieved with adequately high-quality adult-directed speech (ADS) datasets. Reference speech enhancement further improves the observed prosody transfer and the intelligibility of the synthesized speech. Notably, not all IDS features transfer equally well. Temporal characteristics at the syllable level and pitch variability measures show weaker statistical alignment between reference inputs and synthesis outputs, reflecting limitations of the TTS model.

Overall, the findings support the feasibility of synthesizing IDS from naturalistic reference data using reference-based neural TTS. The study highlights both the promise and the current limitations of zero-shot IDS synthesis and provides a foundation for future work aimed at more comprehensive modeling of infant-directed speaking style.

REFERENCES

Act on Early Childhood Education and Care (540/2018) (2018). Section 12: right to early childhood education starts at nine months. URL: https://www.finlex.fi/en/laki/kaannokset/2018/en20180540.pdf.

Ahtinen, A., Beheshtian, N. and Väänänen, K. (2023). Robocamp at home: Exploring families' co-learning with a social robot in the wild. *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 331–340. DOI: 10.1145/3568162.3576976.

Ahtinen, A. and Kaipainen, K. (2020). Learning and Teaching Experiences with a Persuasive Social Robot in Primary School: Findings and Implications from a 4-Month Field Study. *Persuasive Technology – 15th International Conference, PERSUASIVE 2020, Proceedings*. Ed. by S. B. Gram-Hansen, T. S. Jonasen and C. Midden. Vol. 12064. Lecture Notes in Computer Science. Springer, pp. 73–84. DOI: 10.1007/978-3-030-45712-9_6.

Ahtinen, H.-L. A., Chowdhury, A., Beheshtian, N., Ramirez Millan, V. and Wu, C.-H. (2023). Collaborative Learning with Social Robots: Reflections on the Novel Co-learning Concepts Robocamp and Robotour. *Learning and Collaboration Technologies – 10th International Conference, LCT 2023, Proceedings, Part II*. Ed. by P. Zaphiris and A. Ioannou. Vol. 14041. Lecture Notes in Computer Science. Springer, pp. 255–270. DOI: 10.1007/978-3-031-34550-0_18.

Allen, J. B. and Rabiner, L. R. (Nov. 1977). A Unified Approach to Short-Time Fourier Analysis and Synthesis. *Proceedings of the IEEE* 65.11, pp. 1558–1564. DOI: 10.1109/PROC.1977.10770.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F. and Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4218–4222. URL: https://aclanthology.org/2020.lrec-1.520/.

Arik, S., Diamos, G., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J. and Zhou, Y. (2017). *Deep Voice 2: Multi-Speaker Neural Text-to-Speech*. arXiv: 1705.08947 [cs.CL]. URL: https://arxiv.org/abs/1705.08947.

Bäckström, T., Räsänen, O., Zewoudie, A., Pérez Zarazaga, P., Koivusalo, L., Das, S., Gómez Mellado, E., Bouafif Mansali, M., Ramos, D., Kadiri, S., Alku, P. and Vali, M. H. (2022). *Introduction to Speech Processing*. 2nd ed. DOI: 10.5281/zenodo.6821775. URL: https://speechprocessingbook.aalto.fi.

Bahdanau, D., Cho, K. and Bengio, Y. (2015). *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv: 1409.0473 [cs.CL]. URL: https://arxiv.org/abs/1409.0473.

Bänziger, T. and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication* 46.3–4, pp. 252–267. DOI: 10.1016/j.specom.2005.02.016.

Benders, T. and Blom, E. (2023). Computational modelling of language acquisition: an introduction. *Journal of Child Language* 50.6, pp. 1287–1293. DOI: 10.1017/S0305000923000429.

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., Hamrick, L. R., Marklund, E., Kalashnikova, M., Guez, A., Casillas, M., Benetti, L., Alphen, P. van and Cristia, A. (2023). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences* 120.52, e2300671120. DOI: 10.1073/pnas.2300671120.

Betker, J. (2023). *Better speech synthesis through scaling*. arXiv: 2305.07243 [cs.SD]. URL: https://arxiv.org/abs/2305.07243.

Boll, S. F. (1979). Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. Vol. 27. 2, pp. 113–120.

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* 20.3-4, pp. 255–272. DOI: 10.1016/S0167-6393(96)00063-5.

Brent, M. R. (1999). Speech Segmentation and Word Discovery: A Computational Perspective. *Trends in Cognitive Sciences* 3.8, pp. 294–301. DOI: 10.1016/S1364-6613(99)01350-9.

Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S. and Weber, J. (Sept. 2024). XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. *Proceedings of Interspeech 2024*, pp. 4978–4982. DOI: 10.21437/Interspeech.2024-2016.

Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E. and Ponti, M. A. (2022). YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. *Proceedings of the 39th International Conference on Machine Learning*, pp. 2709–2720. URL: https://proceedings.mlr.press/v162/casanova22a.html.

Chawarska, K., Volkmar, F. and Klin, A. (2010). Limited attentional bias for faces in toddlers with autism spectrum disorders. *Archives of General Psychiatry* 67.2, pp. 178–185. DOI: 10.1001/archgenpsychiatry.2009.194.

Chen, L. (2009). Audio Quality Issue for Automatic Speech Assessment. *Proceedings of SLaTE (Speech and Language Technology in Education)*. ISCA. URL: https://www.isca-archive.org.

Close, G., Hain, T. and Goetze, S. (2024). Hallucination in Perceptual Metric-Driven Speech Enhancement Networks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Coqui (2024). *coqui/XTTS-v2: Text-to-Speech Model*. https://huggingface.co/coqui/XTTS-v2. Hugging Face model card.

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G. and Fusaroli, R. (2023). A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour* 7.1, pp. 114–133. DOI: 10.1038/s41562-022-01452-1.

Cychosz, M., Edwards, J. R., Ratner, N. B., Eaton, C. T. and Newman, R. S. (2021). Acoustic-Lexical Characteristics of Child-Directed Speech Between 7 and 24 Months and Their Impact on Toddlers' Phonological Processing. *Frontiers in Psychology* 12, p. 712647. DOI: 10.3389/fpsyg.2021.712647.

Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., Casillas, M., Barbaro, K. de, Bang, J. Y. and Weisleder, A. (2020). Longform recordings of everyday life: Ethics for best practices. *Behavior Research Methods* 52.5, pp. 1951–1969. DOI: 10.3758/s13428-020-01365-9.

Davis, S. and Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4, pp. 357–366. DOI: 10.1109/TASSP.1980.1163420.

De Jong, N. H., Pacilly, J. and Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice* 28.4, pp. 456–476. DOI: 10.1080/0969594X.2021.1951162.

Dellwo, V., Leemann, A. and Kolly, M.-J. (2015). The recognition of read and spontaneous speech in local vernacular: The case of Zurich German. *Journal of Phonetics* 48, pp. 13–28. DOI: 10.1016/j.wocn.2014.10.011.

Demuth, K., Culbertson, J. and Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech* 49.2, pp. 137–174. DOI: 10.1177/00238309060490020201.

Desplanques, B., Thienpondt, J. and Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *Proceedings of INTERSPEECH 2020*. International Speech Communication Association (ISCA), pp. 3830–3834. DOI: 10.21437/Interspeech.2020-2650.

Droucker, D., Curtin, S. and Vouloumanos, A. (2013). Linking infant-directed speech and face preferences to language outcomes in infants at risk for autism spectrum disorder. *Journal of Speech, Language, and Hearing Research* 56.2, pp. 567–576. DOI: 10.1044/1092-4388(2012/11-0266).

ElevenLabs (2025a). *AI Voices Get Emotional: Expressive TTS with Text Cues*. https://datasciencedeveloper.wordpress.com/2025/07/15/ai-voices-get-

emotional-elevenlabs-debuts-its-most-expressive-tts-mode/. Accessed 2025-12-27.

ElevenLabs (2025b). *Capabilities of ElevenLabs Text-to-Speech*. https://elevenlabs.io/docs/overview/capabilities/text-to-speech. Accessed 2025-12-27.

– (2025c). *ElevenLabs Text-to-Speech Models*. https://elevenlabs.io/docs/models. Accessed 2025-12-27.

Englund, N. and Behne, D. M. (2006). Changes in infant directed speech in the first six months. *Infant and Child Development* 15.2, pp. 139–160. DOI: 10.1002/icd.445.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.6, pp. 1109–1121. DOI: 10.1109/TASSP.1984.1164454.

Farran, L. K., Lee, C.-C., Yoo, H. and Oller, D. K. (2016). Cross-Cultural Register Differences in Infant-Directed Speech: An Initial Study. *PLOS ONE* 11.3, e0151518. DOI: 10.1371/journal.pone.0151518.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development* 8.2, pp. 181–195. DOI: 10.1016/S0163-6383(85)80005-9.

Frank, M. C., Alcock, K. J., Arias-Trejo, N., Aschersleben, G., Baldwin, D., Barbu, S. et al. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science* 3.1, pp. 24–52. DOI: 10.1177/2515245919900809. URL: https://journals.sagepub.com/doi/10.1177/2515245919900809.

Fromkin, V., Rodman, R. and Hyams, N. (2018). *An Introduction to Language*. 11th ed. Cengage Learning. ISBN: 9781337559577.

Garellek, M., Samlan, R., Gerratt, B. R. and Kreiman, J. (2016). Modeling the voice source in terms of spectral slopes. *Journal of the Acoustical Society of America* 140.4. DOI: 10.1121/1.4944474.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G. and Pallett, D. S. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium. LDC Catalog No. LDC93S1. URL: https://catalog.ldc.upenn.edu/LDC93S1.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press. URL: https://www.deeplearningbook.org/.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R. and Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems* 28.10, pp. 2222–2232. DOI: 10.1109/TNNLS.2016.2582924.

Grieser, D. L. and Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology* 24.1, pp. 14–20. DOI: 10.1037/0012-1649.24.1.14.

Griffin, D. W. and Lim, J. S. (1984). Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2, pp. 236–243. DOI: 10.1109/TASSP.1984.1164317.

Harris, F. J. (Feb. 1978). On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform. *Proceedings of the IEEE* 66.1, pp. 51–83. DOI: 10.1109/PROC.1978.10837.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). *Deep Residual Learning for Image Recognition*. arXiv: 1512.03385 [cs.CV]. URL: https://arxiv.org/abs/1512.03385.

Hill, D., Manzara, L. and Schock, C. (1995). Real-time articulatory speech-synthesis-by-rules. *Proceedings of AVIOS '95, 14th Annual International Voice Technologies Applications Conference of the American Voice I/O Society* (Sept. 11–14, 1995). San Jose, CA, USA, pp. 27–44. URL: https://cspages.ucalgary.ca/~hilld/papers/avios95/body.htm.

Hillenbrand, J., Cleveland, R. A. and Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research* 37.4, pp. 769–778. DOI: 10.1044/jshr.3704.769.

Ho, J., Jain, A. and Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. arXiv: 2006.11239 [cs.LG]. URL: https://arxiv.org/abs/2006.11239.

Huebner, P. A. and Willits, J. A. (2021). Using lexical context to discover the noun category: Younger children have it easier. *Psychology of Learning and Motivation*. Ed. by B. H. Ross. Vol. 75. Amsterdam: Elsevier, pp. 279–331. DOI: 10.1016/bs.plm.2021.08.002.

Ishikawa, K., Nudelman, C., Park, S. and Ketring, C. (2021). Perception and Acoustic Studies of Vowel Intelligibility in Dysphonic Speech. *Journal of Voice* 35.4, 659.e11–659.e24. DOI: 10.1016/j.jvoice.2019.12.022.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Hénaff, O., Botvinick, M. M., Zisserman, A., Vinyals, O. and Carreira, J. (2022). Perceiver IO: A General Architecture for Structured Inputs & Outputs. *International Conference on Learning Representations (ICLR)*. arXiv:2107.14795.

Jain, A. K., Ross, A. and Prabhakar, S. (2004). An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 14.1, pp. 4–20. DOI: 10.1109/TCSVT.2003.818349.

Jung, J.-W., Tak, H., Shim, H.-J., Heo, H.-S., Lee, B.-J., Chung, S.-W., Yu, H.-J., Evans, N. and Kinnunen, T. (2022). SASV 2022: The First Spoofing-Aware Speaker Verification Challenge. *Proc. Interspeech*, pp. 2893–2897. DOI: 10.21437/Interspeech.2022-11270.

Kim, J., Kim, S., Kong, J. and Yoon, S. (2020). *Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search*. arXiv: 2005.11129 [eess.AS]. URL: https://arxiv.org/abs/2005.11129.

Kinnunen, T. and Li, H. (2010). An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. *Speech Communication* 52.1, pp. 12–40. DOI: 10.1016/j.specom.2009.08.009.

Kong, J., Kim, J. and Bae, J. (2020). *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis*. arXiv: 2010.05646 [cs.SD]. URL: https://arxiv.org/abs/2010.05646.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience* 5.11, pp. 831–843. DOI: 10.1038/nrn1533.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V. and Ryskina, V. L. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science* 277.5326, pp. 684–686. DOI: 10.1126/science.277.5326.684.

Ladefoged, P. and Johnson, K. (2011). *A Course in Phonetics*. 5th ed. Wadsworth. ISBN: 9781428231269.

Lameris, H., Mehta, S., Henter, G. E., Gustafson, J. and Székely, É. (2023). Prosody-Controllable Spontaneous TTS with Neural HMMS. *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10097200.

Lancucki, A. (June 2021). Fastpitch: Parallel Text-to-Speech with Pitch Prediction. *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6588–6592. DOI: 10.1109/ICASSP39728.2021.9413889.

Lavechin, M., Seyssel, M. de, Métais, M., Metze, F., Mohamed, A., Bredin, H., Dupoux, E. and Cristia, A. (2024). Modeling Early Phonetic Acquisition from Child-Centered Audio Data. *Cognition* 244, p. 105734. DOI: 10.1016/j.cognition.2024.105734.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791.

Lee, Y. and Kreiman, J. (2022). Acoustic Voice Variation in Spontaneous Speech. *The Journal of the Acoustical Society of America* 151.5, pp. 3462–3472. DOI: 10.1121/10.0011471.

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M. and Zhou, M. (2019). *Neural Speech Synthesis with Transformer Network*. arXiv: 1809.08895 [cs.CL]. URL: https://arxiv.org/abs/1809.08895.

Lieberman, P. and Blumstein, S. E. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge, UK: Cambridge University Press. ISBN: 9780521313575.

Liu, X., Kong, Q., Zhao, Y., Liu, H., Yuan, Y., Liu, Y., Xia, R., Wang, Y., Plumbley, M. D. and Wang, W. (2023). *Separate Anything You Describe*. arXiv: 2308.05037 [eess.AS]. URL: https://arxiv.org/abs/2308.05037.

Lu, Y.-X., Du, H.-P., Liu, F., Ai, Y. and Ling, Z.-H. (2025). Improving Noise Robustness of LLM-based Zero-shot TTS via Discrete Acoustic Token Denoising. *Interspeech*. DOI: 10.21437/Interspeech.2025-319.

MacWhinney, B. and Snow, C. (1985). The Child Language Data Exchange System. *Journal of Child Language* 12.2, pp. 271–295. DOI: 10.1017/S0305000900006449.

Martin, A., Igarashi, Y., Jincho, N. and Mazuka, R. (2016). Utterances in infant-directed speech are shorter, not slower. *Cognition* 156, pp. 52–59. DOI: 10.1016/j.cognition.2016.07.015.

Masataka, N. (1999). Preference for infant-directed singing in 2-day-old hearing infants of deaf parents. *Developmental Psychology* 35.4, pp. 1001–1005. DOI: 10.1037/0012-1649.35.4.1001.

Menéndez, M. L., Pardo, J. A., Pardo, L. and Pardo, M. C. (1997). The Jensen-Shannon divergence. *Journal of the Franklin Institute* 334.2, pp. 307–318. DOI: 10.1016/S0016-0032(96)00063-4.

Meyer, J., Rauchenstein, L., Eisenberg, J. D. and Howell, N. (2020). Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 6462–6468. URL: https://aclanthology.org/2020.lrec-1.796/.

Miyazawa, K., Shinya, T., Martin, A., Kikuchi, H. and Mazuka, R. (2017). Vowels in infant-directed speech: More breathy and more variable, but not clearer. *Cognition* 166, pp. 84–93. DOI: 10.1016/j.cognition.2017.05.003.

Murray, D. (2018). Monotropism – An Interest Based Account of Autism. *Encyclopedia of Autism Spectrum Disorders*. Cham: Springer. DOI: 10.1007/978-1-4614-6435-8. URL: https://doi.org/10.1007/978-1-4614-6435-8.

Narayan, C. R. and McDermott, L. C. (2016). Speech Rate and Pitch Characteristics of Infant-Directed Speech: Longitudinal and Cross-Linguistic Observations. *The Journal of the Acoustical Society of America* 139.3, pp. 1272–1281. DOI: 10.1121/1.4944634.

Naylor, P. A. (2014). Introduction to Speech Processing. *Academic Press Library in Signal Processing: Volume 4 – Image, Video Processing and Analysis, Hardware, Audio, Acoustic and Speech Processing*. Ed. by J. Trussell, A. Srivastava, A. K. Roy-Chowdhury, A. Srivastava, P. A. Naylor, R. Chellappa and S. Theodoridis. Academic Press. ISBN: 9780123965011.

Odijk, L. and Gillis, S. (2023). Parents tune their vowels to the emergence of children's words. *Journal of Child Language* 50.5, pp. 1184–1203. DOI: 10.1017/S0305000922000289.

Oord, A. van den, 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*. arXiv: 1609.03499 [cs.SD]. URL: https://arxiv.org/abs/1609.03499.

Oord, A. van den, Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G. van den, Lockhart, E., Cobo, L. C., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D. and Hassabis, D. (2017). *Parallel WaveNet: Fast High-Fidelity Speech Synthesis*. arXiv: 1711.10433 [cs.LG]. URL: https://arxiv.org/abs/1711.10433.

Oord, A. van den, Vinyals, O. and Kavukcuoglu, K. (2017). *Neural Discrete Representation Learning*. arXiv: 1711.00937 [cs.LG]. URL: https://arxiv.org/abs/1711.00937.

OpenAI (2023). *GPT-4 Large Language Model*. Accessed via OpenAI ChatGPT interface. URL: https://openai.com/research/gpt-4.

Oxenham, A. J. (2018). How We Hear: The Perception and Neural Coding of Sound. *Annual Review of Psychology* 69, pp. 27–50. DOI: 10.1146/annurev-psych-122216-011635.

Pearson, K. (1895). Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London* 58.347–352, pp. 240–242. DOI: 10.1098/rspl.1895.0041.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J. and Miller, J. (2018). Deep Voice 3: 2000-Speaker Neural Text-to-Speech. *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HJtEm4p6Z.

Quintero, N. M. and Ahtinen, A. (2023). University Language Instructors Programming Robotic Learning Applications: Design and Implementation of Encouraging Programming Workshop Experiences. *Proceedings of the 2023 Conference on Learning at Scale (L@S)*. DOI: 10.1145/3623809.3623870.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* 77.2, pp. 257–286. DOI: 10.1109/5.18626.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. et al. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv: 2212.04356 [eess.AS]. URL: https://arxiv.org/abs/2212.04356.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. OpenAI Technical Report. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Communication* 54.9, pp. 975–997. DOI: 10.1016/j.specom.2012.05.001.

Räsänen, O., Cruz Blandón, M., Khorrami, K. and Kocharov, D. (Apr. 2024). Modeling Child Language Development using Naturalistic Data at a Scale. *Finnic Phonetics Symposium*. Tallinn, Estonia. URL: https://cs.ttu.ee/events/fp2024/FP2.

Räsänen, O., Doyle, G. and Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition* 171, pp. 130–150. DOI: 10.1016/j.cognition.2017.11.003.

Räsänen, O. and Kocharov, D. (2025). A pipeline for stochastic and controlled generation of realistic language input for simulating infant language acquisition. *Behavior Research Methods* 57.275. DOI: 10.3758/s13428-025-02750-0.

Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation), Article 4(14) (2016). Defines biometric data as "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person." URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation), Article 9(1) (2016). Article 9(1) states that the processing of special categories of personal data, including biometric data used for the purpose of uniquely identifying a natural person, is prohibited unless specific conditions defined in the regulation apply. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y. (2019). *FastSpeech: Fast, Robust and Controllable Text to Speech*. arXiv: 1905.09263 [cs.CL]. URL: https://arxiv.org/abs/1905.09263.

ResembleAI (2023). *Introducing Resemble Enhance: Open Source Speech Super Resolution AI Model*. URL: https://www.resemble.ai/introducing-resemble-enhance.

Roseberry, S., Hirsh-Pasek, K. and Golinkoff, R. M. (2014). Skype Me! Socially Contingent Interactions Help Toddlers Learn Language. *Child Development* 85.3, pp. 956–970. DOI: 10.1111/cdev.12166.

Samposalo, H., Puura, K., Kaipainen, K., Jyskä, I., Bagaskara, I., Tulaskar, R. and Turunen, M. (2024). Social robotics in promoting children's positive emotions and reciprocity in early childhood education: Gamified simulation as a data collection tool. *WAIMH 2024 Interim World Congress on Infant Mental Health: Abstracts*. Conference abstract. Tampere, Finland.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y. and Wu, Y. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783. DOI: 10.1109/ICASSP.2018.8461368.

Shin, Y., Lee, Y., Jo, S., Hwang, Y. and Kim, T. (2022). Text-driven Emotional Style Control and Cross-speaker Style Transfer in Neural TTS. *Proceedings of Interspeech*. DOI: 10.21437/Interspeech.2022-10131.

Shinozaki, T., Ostendorf, M. and Atlas, L. (2009). Characteristics of Speaking Style and Implications for Speech Recognition. *The Journal of the Acoustical Society of America* 126.3, pp. 1500–1510. DOI: 10.1121/1.3183593.

Silverman, K. E. A., Blaauw, E., Spitz, J. and Pitrelli, J. F. (1992). A Prosodic Comparison of Spontaneous Speech and Read Speech. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. International Speech Communication Association (ISCA), pp. 1299–1302. URL: https://www.isca-archive.org/icslp_1992/silverman92b_icslp.html.

Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R. and Saurous, R. A. (2018). *Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron*. arXiv: 1803.09047 [cs.CL]. URL: https://arxiv.org/abs/1803.09047.

Snow, C. E. (1972). Mothers' Speech to Children Learning Language. *Child Development* 43.2, pp. 549–565. DOI: 10.2307/1127555.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review* 27.4, pp. 501–532. DOI: 10.1016/j.dr.2007.06.002.

Song, J. Y., Demuth, K. and Morgan, J. (2010). Effects of the Acoustic Properties of Infant-Directed Speech on Infant Word Recognition. *The Journal of the Acoustical Society of America* 128.1, pp. 389–400. DOI: 10.1121/1.3419786.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E. H. and Frank, M. C. (2021). SAYCam: A Large, Longitudinal Audiovisual Dataset Recorded From the Infant's Perspective. *Open Mind: Discoveries in Cognitive Science* 5, pp. 20–29. DOI: 10.1162/opmi_a_00039.

Sutskever, I., Vinyals, O. and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*. Vol. 27. DOI: 10.5555/2969033.2969173.

Tachibana, H., Uenoyama, K. and Aihara, S. (2018). Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4784–4788. DOI: 10.1109/ICASSP.2018.8461829.

Taylor, P. (2009). *Text-to-Speech Synthesis*. 1st. Cambridge, United Kingdom: Cambridge University Press. ISBN: 978-0-521-11692-2.

Themistocleous, C. (2025). Linguistic and Emotional Prosody: A Systematic Review and ALE Meta-Analysis. *Neuroscience & Biobehavioral Reviews* 175, p. 106210. DOI: 10.1016/j.neubiorev.2025.106210.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T. (2000). Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vol. 3. IEEE, pp. 1315–1318. DOI: 10.1109/ICASSP.2000.861820.

Tom, K., Titze, I. R., Hoffman, E. A. and Story, B. H. (2001). Three-dimensional vocal tract imaging and formant structure: Varying vocal register, pitch, and loudness. *The Journal of the Acoustical Society of America* 109.2, pp. 742–747. DOI: 10.1121/1.1332380.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J. and Dupoux, E. (2021). VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 993–1003. DOI: 10.18653/v1/2021.acl-long.80. URL: https://aclanthology.org/2021.acl-long.80/.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S. and Wei, F. (2023). *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers*. arXiv: 2301.02111 [cs.CL]. URL: https://arxiv.org/abs/2301.02111.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R. and Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. *Interspeech 2017*, pp. 4006–4010. DOI: 10.21437/Interspeech.2017-1452.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y. and Saurous, R. A. (2018). *Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis*. arXiv: 1803.09017 [cs.CL]. URL: https://arxiv.org/abs/1803.09017.

Williams, J. and King, S. (2019). Disentangling Style Factors from Speaker Representations. *Interspeech*. URL: https://www.isca-archive.org/interspeech_2019/williams19c_interspeech.html.

Xin, D., Saito, Y., Takamichi, S., Koriyama, T. and Saruwatari, H. (2021). Cross-lingual Speaker Adaptation using Domain Adaptation and Speaker Consistency Loss for Text-To-Speech Synthesis. *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, pp. 1614–1618. DOI: 10.21437/Interspeech.2021-897.

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. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOI: 10.1109/ICASSP40776.2020.9053795.

Yang, D., Liu, S., Huang, R., Weng, C. and Meng, H. (2023). *InstructTTS: Modelling Expressive TTS in Discrete Latent Space with Natural Language Style Prompt*. arXiv: 2301.13662 [cs.SD]. URL: https://arxiv.org/abs/2301.13662.

Yu, F. and Koltun, V. (2016). *Multi-Scale Context Aggregation by Dilated Convolutions*. arXiv: 1511.07122 [cs.CV]. URL: https://arxiv.org/abs/1511.07122.

Zahner, K., Schönhuber, M., Grijzenhout, J. and Braun, B. (2016). Konstanz prosodically annotated infant-directed speech corpus (KIDS Corpus). *Proceedings of the 8th Inter-*

national Conference on Speech Prosody. Boston, USA. URL: https://www.isca-archive.org/speechprosody_2016/zahner16_speechprosody.html.

Zen, H., Senior, A. and Schuster, M. (2013). Statistical Parametric Speech Synthesis Using Deep Neural Networks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOI: 10.1109/ICASSP.2013.6639215.

Zhang, C., Ren, Y., Tan, X., Liu, J., Zhang, K., Qin, T., Zhao, S. and Liu, T.-Y. (2020). *DenoiSpeech: Denoising Text to Speech with Frame-Level Noise Modeling*. arXiv: 2012.09547 [eess.AS]. URL: https://arxiv.org/abs/2012.09547.

Zhang, L., Ren, Y., Deng, L. and Zhao, Z. (2023). HiFiDenoise: High-Fidelity Denoising Text to Speech with Adversarial Networks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Zhou, X., Wang, L., Hong, X. and Wong, P. C. M. (2024). Infant-directed speech facilitates word learning through attentional mechanisms: An fNIRS study of toddlers. *Developmental Science* 27.1, e13424. DOI: 10.1111/desc.13424.

Zhu, Q., Gu, Y., Chen, R., Weng, C., Hu, Y., Dai, L. and Zhang, J. (2023). *Rep2wav: Noise Robust Text-to-Speech Using Self-Supervised Representations*. arXiv: 2308.14553 [eess.AS]. URL: https://arxiv.org/abs/2308.14553.