University of Silesia in Katowice
Faculty of Humanities
Institute of Linguistics

Sapienza University of Rome
Department of European, American,
and Intercultural Studies

Steven Jarosz

A NEURAL NETWORK APPROACH TO MODELING THE ACOUSTIC EFFECTS OF PHONOLOGICAL NEIGHBORHOOD DENSITY AND WORD FREQUENCY IN POLISH SPEECH

Doctoral thesis written
under the supervision of
prof. dr hab. Arkadiusz Rojczyk
and
Prof. Margherita Dore

2026

Table of Contents

Introduction

Speech production is not mechanically uniform. When speakers produce words, their articulation is shaped by the psycholinguistic properties of those words — how frequently they occur in the language, the demands placed on the speaker during lexical access, and how many phonologically similar words exist in the mental lexicon. Words that are harder to access, whether because they are infrequent or because they compete with other phonologically similar neighbors, tend to elicit more precise articulation (hyperarticulation). Words that are easier to access tend to be produced with less precision (hypoarticulation).

The relationship between lexical properties and acoustic output reflects the interactive nature of phonetic encoding and lexical access, and has been extensively documented in production and perception research. This phenomenon is referred to in this thesis as systematic articulatory variability (SAV). SAV sits at the intersection of phonetics and psycholinguistics, and provides insight into how lexical properties shape speech, and whether AI speech systems reproduce those properties.

SAV emerges in the vowel space from the effects of phonological neighborhood density (PND) and word frequency (WF). Low-frequency words in dense neighborhoods are typically produced with more exaggerated vowels — a hyperarticulation strategy that facilitates comprehension. Frequent words in sparse neighborhoods tend to be hypoarticulated, reflecting articulatory economy.

The psycholinguistic literature on PND and WF has been concentrated on English (e.g., Luce & Pisoni, 1998; Vitevitch & Luce, 1998; Munson & Solomon, 2004; Balota et al., 2007) and other widely studied languages (e.g., Scarborough et al., 2018 for French; Tomaschek et al., 2013 for German) — a pattern consistent with the broader English-centric tendencies in cognitive science (Blasi et al., 2022) and in psycholinguistics (Berghoff & Bylund, 2025). This pattern, in which low-frequency words in high-density neighborhoods are produced with hyperarticulation, and vice versa, is well-attested and has been foundational to models of production and perception, including the Hypo- and Hyperarticulation (H&H) theory (Lindblom, 1990), the Neighborhood Activation Model (NAM) (Luce & Pisoni, 1998), and usage-based phonology (Bybee, 2001).

Less-studied languages have received comparatively little attention. Systematic cross-linguistic research on such languages regarding PND and WF remains limited. Polish provides a particularly useful case for investigation, as its vowel inventory is smaller than in English and more dispersed. The degree to which findings from English generalize to Polish remains largely untested. Establishing whether SAV operates similarly in Polish is therefore an empirical and theoretical contribution to the question of how universal lexical effects on phonetic production emerge.

A separate but related question concerns artificial intelligence. Speech synthesis and speech recognition systems are ubiquitous, yet the extent to which they reproduce and perceive the psycholinguistic properties of human speech is not fully understood. If SAV results from human factors such as lexical access and articulatory economy during production, it is not obvious that AI systems would reproduce it. However, since such systems are trained on large corpora of human speech, they may implicitly encode aspects of SAV. With the exception of a single study on speech synthesis and speech recognition in English (Song et al., 2025), the ability of modern AI systems to reproduce SAV has not been empirically tested.

Systems for both speech synthesis and speech recognition have improved dramatically in perceived naturalness and accuracy. However, naturalness and accuracy do not completely characterize what a speech system has learned. Human speech production is determined by cognitive and psycholinguistic processes that leave systematic effects in the acoustic signal. The question of whether AI systems reproduce those effects is largely unexamined. A system may produce acoustically plausible vowels without encoding the lexical sensitivity that embodies the SAV of human speakers.

A methodological contribution of this thesis concerns the approach taken in speech recognition analysis. Rather than evaluating speech recognition accuracy as an output measure, this study considers the internal activations and confidence metrics of Meta's wav2vec 2.0 model. This method reveals how the effects of SAV are encoded — or fail to be encoded — in a state-of-the-art transformer speech recognition model. This thesis compares two drivers of SAV — PND and WF — in Polish human and AI speech production, and examines whether current AI recognition systems exhibit comparable patterns.

The thesis makes three contributions. First, it extends the empirical base for SAV beyond English to Polish, a language that has received relatively little attention in the psycholinguistic

literature on lexical effects in phonetic production. Second, it tests whether commercial AI synthesis models reproduce SAV in a non-English language, addressing a gap that currently rests on a single study of English (Song et al., 2025). Third, it adapts two methods from prior work (tom Dieck et al., 2022; Ravuri et al., 2024) for probing speech recognition models in the Polish language via their internal activations and confidence metrics.

The research questions for this study are as follows:

1. **Human speech data:** How do PND and WF influence the acoustic realization of vowels in Polish? Specifically, are words in dense phonological neighborhoods with low lexical frequency ('hard' words) characterized by hyperarticulation, while words in sparse neighborhoods with high lexical frequency ('easy' words) show hypoarticulation?
2. **AI speech data:** In Polish, to what extent do transformer-based speech synthesis models reproduce the PND- and WF-mediated effects found in the human vowel space?
3. **AI recognition data:** In Polish, do the internal representations of words within a transformer-based speech recognition model reflect a structure analogous to human vowel space? In particular, do 'hard' and 'easy' words occupy distinct regions within the model's representational space, corresponding with their acoustic differentiation in human speech, and do confidence metrics vary by difficulty condition like human perception?

This study is limited in scope. It focuses on PND and WF as drivers of variation in vowel space in Polish. Other acoustic dimensions, including vowel duration and voice onset time, are not examined, and the stimuli are selected from a controlled word list designed to isolate PND and WF effects. On the AI side, the study examines two commercial speech synthesis systems — Amazon's Polly and ElevenLabs — and one recognition model, Meta's wav2vec 2.0. These were selected as representative of current speech technology. The synthesis models are effectively black boxes, as their internal architectures and training data are not publicly available, which constrains the interpretations that can be offered. Although the wav2vec 2.0 model is open source, this analysis is limited to a subset of its activations. Finally, the study is centered on Polish; the findings do not necessarily generalize to other typologically similar languages.

The thesis is organized into six chapters. Chapter 1 establishes the relevant psycholinguistic and phonetic background for this study. It reviews the effects of PND and WF on speech production and perception, describes the acoustic properties of Polish vowels, and presents the

theoretical frameworks most relevant to SAV. Chapter 2 presents an overview of the AI systems examined in this study, introducing the architectures of modern speech synthesis and recognition, and situates them within the broader history of computational approaches to spoken language. The study's methodology is discussed in Chapter 3. It covers the design of the stimulus set, data collection, acoustic measurements, and the statistical approaches used for human and AI data. Chapter 4 presents the results across the three components of the study: SAV effects in Polish human production, their absence in synthetic speech, and a distinct pattern of model confidence in speech recognition. Chapter 5 interprets the findings using the frameworks from Chapters 1 and 2, compares them with previous work on SAV and AI speech systems, and proposes explanations for the observed patterns. Chapter 6 addresses practical applications of the findings, including implications for the development of more naturalistic speech synthesis, the detection of synthetic speech, and the use of SAV-based techniques in second language learning contexts.

Chapter 1: Systematic Articulatory Variability and Polish Vowels

This chapter surveys research on how psycholinguistic factors shape vowel space and acoustic properties, as well as key phonetic aspects of Polish. The coverage of Polish phonetics includes the formant structure and dynamic qualities of vowels, and their roles in both production and perception. Furthermore, the chapter examines psycholinguistic research into the effects of phonological neighborhood density (PND) and word frequency (WF). While well-documented in more widely studied languages such as English (Luce & Pisoni, 1998; Munson & Solomon, 2004; Scarborough & Zellou, 2013; Stephenson, 2004; Vitevitch & Luce, 1998), German (Tomaschek et al., 2013), and French (Scarborough et al., 2018), these effects have received scant attention in underrepresented languages such as Polish. This chapter situates the nascent body of work on Polish within this ongoing debate and provides a foundation for the rest of the thesis.

A notable gap in the literature concerns the interaction between phonetics and lexical properties such as frequency and neighborhood density. In English, high neighborhood density and low frequency are known to elicit hyperarticulation, with expanded vowel spaces and longer vowel durations (Lindblom, 1990; Luce & Pisoni, 1998; Wright, 2004). These effects are in line with usage-based accounts of phonology (Bybee, 2001; Pierrehumbert, 2001). Prior work on Polish includes surprisal effects on vowel dispersion (Malisz et al., 2018) and PND and WF effects on fricative duration (Kaźmierski, 2019). PND and WF effects on Polish vowel production and perception remain largely unexamined.

This chapter establishes the acoustic profile of Polish consonants and vowels, considering both formant targets and their dynamic properties. It also places Polish within theoretical debates on PND and WF, comparing evidence with patterns observed in other languages. A recurring claim is that Polish vowels are "pure in quality," in contrast to the diphthongized monophthongs of English, which display marked formant trajectories over the duration of the vowel (Schwartz, 2021). This is supported by evidence that suggests Polish vowels exhibit reduced formant movement compared to English, particularly in F1 during the first two intervals of the vowel (Schwartz, 2021). However, only a handful of empirical studies have tested this claim. Whether this stability stems only from phonological encoding, as argued in Schwartz's (2010) Onset

Prominence Framework, or whether methodological differences are a confounding factor, remains unresolved. These dynamics may play a role in the effects of PND and WF in Polish.

Equally significant are the research gaps concerning the role of WF in shaping Polish phonetics. In English, low-frequency words are consistently associated with greater articulatory effort, including expanded vowel spaces and longer segment durations, as speakers compensate for their reduced lexical accessibility (Lindblom, 1990; Umeda, 1975; Wright, 2004). These findings have been central to models of lexical access and speech production, reinforcing the idea that gradient frequency effects are encoded in phonetic realizations (Bybee, 2001; Pierrehumbert, 2001). In Polish, however, systematic evidence for frequency-driven phonetic variation remains sparse. This absence of clear frequency effects in Polish reveals an important research gap, raising the possibility that the relationship between lexical frequency and phonetic realization may be language specific.

This chapter provides a background on how phonetic and lexical factors in Polish shape both speech production and perception, setting the stage for the empirical investigations that follow.

1.1. Conceptual Linguistic Framework

1.1.1. Generative Phonology

Since Chomsky & Halle (1968), theoretical linguistics has largely assumed that speech can be decomposed into a finite inventory of discrete phonetic or phonemic units, each functioning as an abstract psychological representation. Generativist linguistics treats language as a symbolic, rule-governed system — one largely independent of usage, frequency, or processing demands (Chomsky, 1965). Lexical representations are insulated from surface variability. Port (2010) argues that empirical data from speech production and perception do not support the generative concept of segmental boundaries. In this view, generative phonology has paid relatively little attention to the phonetic richness now considered central to understanding how speech is acquired and used.

1.1.2. Neighborhood Activation Model (NAM)

Luce and Pisoni's (1998) Neighborhood Activation Model (NAM) is a framework for understanding lexical and sublexical effects on spoken word recognition. At the lexical level, a set of phonologically similar neighbors is activated in memory when a spoken word is heard. This activation creates competition among the neighbors, with the highest-activation word winning. Activation levels are determined by both a word's phonological similarity to other words in the lexicon and its frequency, with higher-frequency words receiving stronger activation. NAM predicts that words from high-density neighborhoods (large sets of phonologically similar words) will be recognized more slowly and less accurately than words from low-density neighborhoods.

A secondary prediction is that lexical effects are muted when sublexical effects are present. For example, Luce & Pisoni (1998) found that experimental nonwords from high-density neighborhoods with a high phonotactic probability (i.e., frequently occurring bi-phone sequences in the language) are recognized faster. From a lexical point of view, the expected effect would be slower recognition; however, non-words do not exist in the lexicon and are not subject to lexical effects. Therefore, Luce & Pisoni (1998) suggest that two levels of competition exist, in which higher neighborhood density (lexical-level) results in inhibition, while higher phonotactic probability (sublexical-level) results in facilitation.

1.1.3. Vowel Inherent Spectral Change (VISC)

Nearey and Assmann (1986) introduced the term Vowel Inherent Spectral Change (VISC) to describe the dynamic formant changes that occur over the course of the vowel. In this framework, such changes in spectral properties are central to vowel perception. The dynamic approach to vowels contrasts with the traditional, and still prevalent, notion of static vowel targets with steady-state formants.

1.1.4. Silent Center

In response to the broader shift toward a dynamic view of vowels, Strange et al. (1983) proposed the Silent Center paradigm. Strange et al. (1983) removed the central, stable portion of a vowel, leaving only its onset and offset transitions, to test if subjects could still reliably identify the vowel. They showed that listeners recognized the vowels correctly despite the removal of the

steady-state portion, or the “silent center.” This research supported the dynamic interpretation of vowels embodied in the VISC paradigm.

1.1.5. The Concept of Vowel Space

Vowel space is a fundamental tool in acoustic phonetics: a two-dimensional plot of the first two vowel formants (F1 and F2) that visualizes a speaker's vowel system. The frequencies of the first two formants of vowels (F1 and F2) are plotted on a two-dimensional plot. As these formants reflect the resonant frequencies of the vocal tract, F1 correlates with tongue height or the openness of the vocal tract, while F2 correlates with tongue advancement (i.e., front vs. back vowels).

Traditionally, a vowel system is depicted with discrete points corresponding to individual phonemes. A speaker’s vowel space is displayed by plotting vowels using F1 on the y-axis and F2 on the x-axis. The shape of the vowel space is delimited by the formants of the most peripheral vowels: front, back, high, and low. The vowel space area can be delineated using various geometric methods, including simple polygons (e.g., triangle or quadrilateral), convex hull (which includes all perimeter vowels and maximizes the area), or concave hull (which provides a more conservative space and carries more local information). Figure 1.1 depicts the Polish vowel space.

.

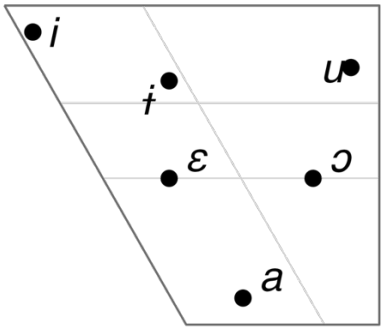

Figure 1.1. Polish vowel chart based on formant values from Kudela-Dobrogowska (1973), as reported in Jassem (1992).

Different procedures exist for obtaining these formant values, which can meaningfully impact the resulting vowel space. The most straightforward and commonly used method involves

extracting formant values at the temporal midpoint of each vowel. Another common approach calculates the mean formant values from the middle third of the vowel's duration. Calculating the point of minimal movement in formant trajectories offers an alternative approach. This method approximates the vowel's steady-state target, using the segment typically found between 20% and 80% of the vowel's duration.

Despite its many advantages — such as versatility and the convenience of visualization — the vowel space system has its limitations. Traditional vowel space measures rely on a static point of the F1 and F2 formants at a single point in time in the duration of the vowel. This omits dynamic spectral properties such as formant trajectories. This is especially problematic for diphthongs, which exhibit continuous movement through the acoustic space, and for nasal vowels (e.g., Polish ą and ę) whose recognition relies on more than F1 and F2. Though the F1–F2 representation may oversimplify vowel identity by omitting higher formants and formant dynamics, it remains the primary method for studying vowels in acoustic phonetics.

1.2. Acoustic Properties of Polish Vowels

Polish features a relatively compact vocalic system, composed of six oral monophthongs: /i, ɨ, e, a, ɔ, u/ and two nasal vowels, /ɔ̃/ and /ɛ̃/. This comparatively small inventory (smaller than, say, Germanic languages) results in a "sparsely filled" vowel space.

1.2.1. Dynamic Changes in Polish Vowels

While formant trajectories vary significantly across languages, their role in less-studied languages has not been extensively explored. In contrast to English, most existing phonetic descriptions of Polish make no explicit mention of formant trajectories. A notable exception is Schwartz (2021), who demonstrated that British English exhibits a greater degree of F1 movement in its vowel system than Polish, with this movement being concentrated earlier in the vowel's duration. Schwartz reports that Polish shows more F1 movement than English in the third interval of the vowel, whereas English shows greater F1 movement overall and earlier in the vowel.

Though both F1 and F2 trajectories are affected by onset consonants, the phonologization of formant trajectories is more consistently observable in F1 than in F2 (Schwartz, 2021). This is because the F1 effect is more consistent over various points of articulation — for example, labial,

coronal, and velar stops all tend to induce a rising F1 trajectory. In contrast, F2 trajectories vary based on the consonant's point of articulation. According to Schwartz, the greater formant movement in Polish, in particular F1, occurs in the third interval of the vowel, by contrast to English. This may provide a rationale for the so-called pureness of Polish vowels. To capture vowel features, traditional vowel-space methods that examine a single point along the vowel trajectory are insufficient. Vowel Inherent Spectral Change (VISC) (Nearey & Assmann, 1986) provides a method for doing so by plotting multiple points of the F1 and F2 trajectories over the duration of the vowel.

1.3. Phonological Neighborhood Density (PND)

Phonological neighborhood density (PND) is the number of words in the lexicon that differ from a target word by a single phoneme — through addition, deletion, or substitution (Luce & Pisoni, 1998; Vitevitch & Luce, 1999). While PND remains an underdeveloped area of research, existing studies show that words from dense neighborhoods are both produced and perceived differently than those from sparse neighborhoods (Luce & Pisoni, 1998; Vitevitch & Luce, 1999). High-density words are embedded in a set of many lexical competitors, whereas low-density words have less competition. This distribution has direct consequences for spoken word recognition: high-density items are usually recognized more slowly and less accurately due to greater competition, while low-density items benefit from reduced lexical interference (Luce & Pisoni, 1998).

The density of the neighborhood is calculated and represented by the PND value. A higher count indicates a denser neighborhood, while a lower count indicates a sparser neighborhood. The most traditional way to define a word's neighborhood is within the distance of a single-phoneme "edit" — a calculation consisting of a single phonological change when one word differs from another word by just one phoneme (see Figure 1.1).

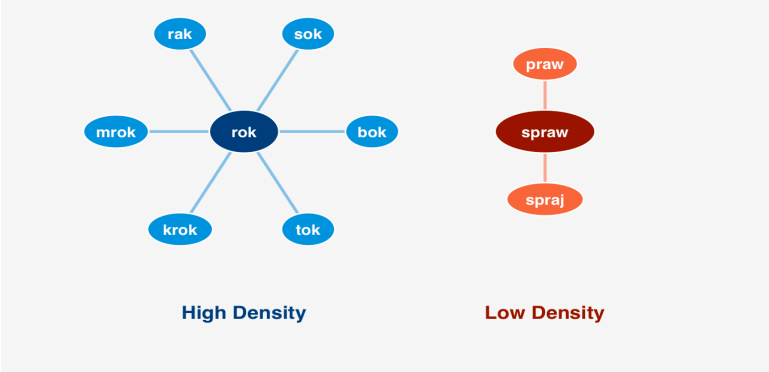

Figure 1.2. Example of high-density and low-density neighborhoods in Polish based on a single-phoneme edit (author-created).

This dissimilarity measurement, known as the "edit distance" or "Levenshtein distance," quantifies the minimum number of single-phoneme alterations — substitution, addition, or deletion — required to transform one word into another (Levenshtein, 1966). Levenshtein distance was originally developed in information theory and is now used widely in computational linguistics, natural language processing, and speech recognition (Sanders & Chin, 2009). These metric measures phonological or orthographic similarity, and is standardly used to quantify lexical proximity, identify variants, and model perceptual discriminability (Yarkoni et al., 2008). When applied to phoneme sequences, the Levenshtein distance enables fine-grained comparisons of word forms based on segmental structure, a method relevant to psycholinguistic modeling and corpus phonology (Nerbonne & Heeringa, 1997).

Building on the Levenshtein distance is the PLD20, Phonological Levenshtein Distance 20. PLD20 is the mean Levenshtein distance from a target word to its 20 phonologically closest neighbors in the lexicon (Alzahrani, 2025). Consequently, a lower PLD20 value for a target word signifies a greater number of highly similar phonological neighbors in the mental lexicon (Alzahrani, 2025). Figure 1.3 illustrates how PLD20 is calculated for a given target word.

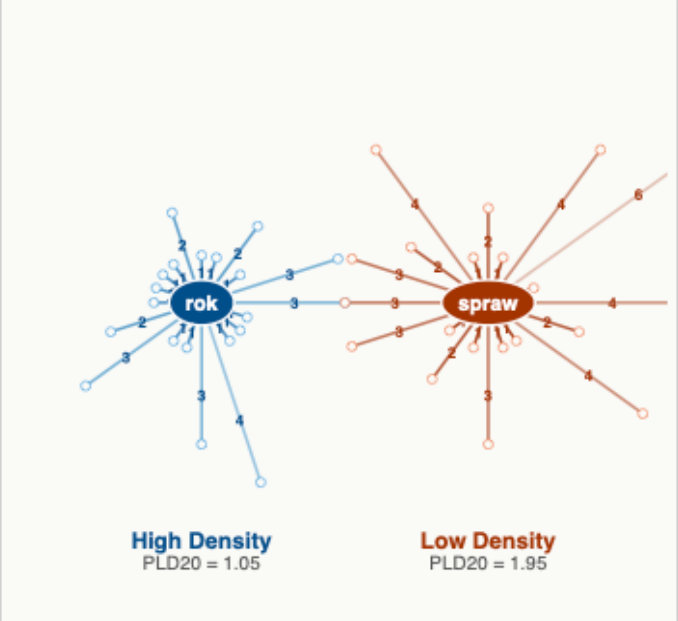

Figure 1.3. Example of PLD20 in Polish (author-created).

Another method used to calculate Levenshtein distance is the network-based "two-hop density," which measures both one-edit and two-edit phonological neighbors (Siew & Vitevitch, 2019). Two-hop density extends this to second-order neighbors — the neighbors of each intermediate neighbor — providing a measure of extended neighborhood structure, as shown in Figure 1.4. A related measure is the clustering coefficient, which quantifies the extent to which a word's immediate phonological neighbors are interconnected. Both measures test whether lexical competition is affected by the extended phonological neighborhood (Alzahrani, 2025).

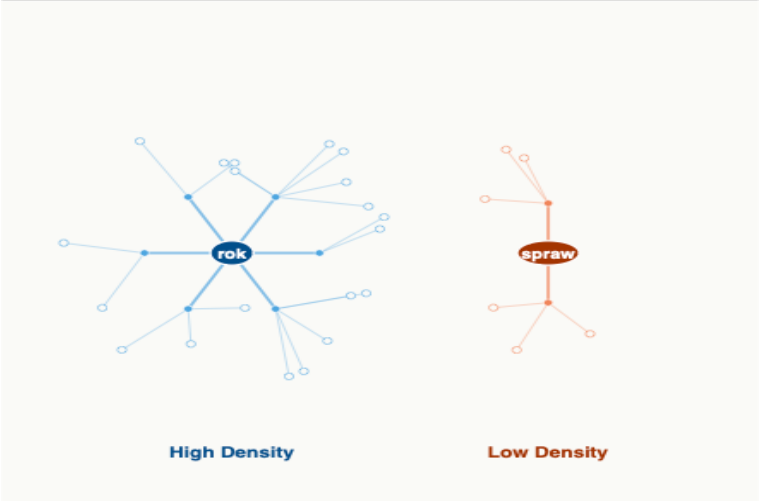

Figure 1.4. Example of two-hop phonological neighborhood density in Polish (author-created).

1.3.1. Neighborhood Density Effects on Speech Perception

Building on early models of lexical access proposed by Marslen-Wilson & Welsh (1978) and McClelland & Elman (1986), contemporary theories of lexical competition in word recognition, including NAM, are based on a model in which spoken words are recognized through competition among phonologically similar neighbors (Luce & Pisoni, 1998). Although this is a simplification, the core tenet of NAM is that words are recognized more slowly and less accurately as the number of phonologically similar neighbors increases (Luce & Pisoni, 1998).

Evidence for neighborhood density effects has been found in English (Luce & Pisoni, 1998; Vitevitch & Luce, 1998), French (Dufour & Frauenfelder, 2010; Ziegler et al., 2003), and Spanish (Vitevitch & Rodríguez, 2005). This inhibitory effect on reaction times in spoken-word recognition is typically attributed to heightened lexical competition among activated phonological neighbors within NAM (Luce & Pisoni, 1998). Although high PND typically slows and reduces the accuracy of auditory word recognition through lexical competition, contradictory findings have been reported. Vitevitch & Rodríguez reported a facilitative PND effect in Spanish auditory word recognition, in which words from dense neighborhoods were recognized faster. Language-specific factors have been proposed to account for this difference. In Spanish, for example, phonological neighbors frequently overlap with morphological neighbors, producing additional activation (Vitevitch & Rodríguez, 2005). A similar facilitation effect has been reported for Russian, where words in dense phonological neighborhoods are recognized faster than those in sparse ones

(Arutiunian & Lopukhina, 2020). Vitevitch and Rodríguez, in line with Vitevitch & Stamer (2006), posit that the morphological and highly inflected nature of Russian is the driver of this effect. These results suggest that other highly inflected languages, such as Polish, may exhibit similar patterns.

The impact of PND appears to be task dependent. In visual word recognition tasks, high PND typically facilitates processing, leading to faster response times (Yates et al., 2004). Because visual recognition does not require phonological processing, this finding that phonological representations are implicitly activated even when not explicitly required by the task. However, in bilingual populations, cross-linguistic phonological similarity can induce inhibitory effects on response times, suggesting competition between phonological representations across languages (Dijkstra et al., 1999).

While PND has been studied for decades, data on its possible effects on recent AI models for speech recognition and speech synthesis are all but nonexistent. As of this writing, only one study has been published in this nascent field. In contrast to the well-documented effects of PND and WF in human speech, Song et al. (2025) found that PND did not play a significant role in computer speech recognition — at least, not under normal speaking conditions. Amazon's Alexa application was tested for its ability to recognize (human) participants' production of $C^{+}VC^{+}$ target words at two speaking rates (normal and fast). No significant effect was found in Alexa's recognition of words under normal speaking conditions with regard to PND, suggesting that Alexa does not have an internal representation that simulates lexical competition in the same way as humans. However, a statistically significant effect was found for low PND words, which were recognized with greater accuracy than high-density words when produced at a fast speaking rate.

Song et al. (2025) suggest that such effects may be "dormant" under normal conditions (i.e., a normal speaking rate) but may become exaggerated under more challenging conditions, such as a fast speaking rate. Overall, their results do not provide substantial support for meaningful PND effects in Alexa, and caution should be taken when extrapolating and even anthropomorphizing an AI model by suggesting that these effects could result from a similar mental representation to that in humans. As will be discussed later, AI models do not learn language in the same way as humans, neither in their approach to acquiring language competence nor in the contents of their training data, and an AI language model cannot be assumed to be a faithful representation of the language model in humans.

1.3.2. Neighborhood Density Effects on Speech Production

Phonological neighborhood density also has notable effects in speech production. The density of a word's phonological neighborhood has been shown to prompt speakers to make measurable phonetic adjustments in the vowel space, durational properties, and consonantal articulation. In dense phonological neighborhoods, where lexical competition is strongest, speakers tend to hyperarticulate target words to maintain their perceptual distinctiveness (Munson & Solomon, 2004; Scarborough & Zellou, 2013). These articulatory modifications could be a speaker's unconscious attempt to enhance the distinctiveness of words that might otherwise slow word recognition due to increased lexical competition from numerous phonological neighbors.

Such an interpretation aligns with theories positing a balance between listener-oriented clarity and speaker-oriented articulatory effort in speech planning. Lindblom's (1990) Hypo- and Hyperarticulation (H&H) theory posits a speaker's ongoing effort to balance communicative clarity for the listener with articulatory economy. Hypoarticulation is produced by means of articulatory undershoot, in which the vowel target is closer to the Euclidean center of the vowel space, resulting in a contracted vowel space. In contrast, hyperarticulation is achieved with overshoot by producing vowels further from the Euclidean center, resulting in an expanded vowel space. This expansion and contraction of the vowel space is usually measured as the Euclidean distance in F1–F2 (Bark-transformed) space (Stephenson, 2004; Munson & Solomon, 2004). By producing the vowels of high-PND words more peripherally, speakers increase the acoustic distance, or contrast, between vowels, enhancing a word's discriminability from its neighbors. The acoustic consequences of hypo- and hyperarticulation are displayed in Figure 1.4.

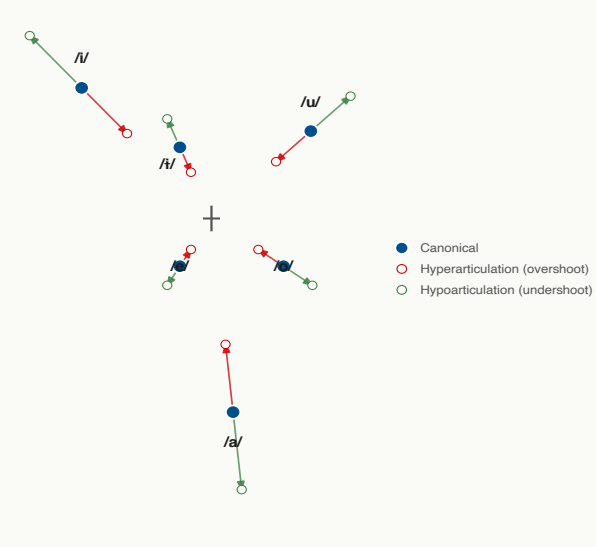

Figure 1.4. Hypo- and Hyperarticulation (H&H) in Polish vowel space (author-created).

Similar patterns have been documented cross-linguistically, albeit with language-specific nuances. In French, oral vowels in high-density neighborhoods are produced more peripherally than those in low-density neighborhoods, paralleling the English pattern (Scarborough et al., 2018). Scarborough et al. found a reversal of the hyperarticulation effect: nasal vowels in high-ND words were hypoarticulated (contracted in vowel space) relative to those in low-PND words. Scarborough et al. interpreted this contraction as an enhancement strategy, leveraging the natural tendency of nasal vowels to be centralized (especially in F1) compared to their oral counterparts. By further centralizing nasal vowels, speakers increase their pairwise contrast with oral vowels, enhancing their distinctiveness in combination with nasalization.

Additionally, Stephenson (2004) reports the opposite of the expected neighborhood density effect, finding that greater vowel-space expansion occurs for nonwords from sparse neighborhoods. One interpretation is that the absence of stored lexical representations alters how articulatory effort is applied during nonword production. Producing an unfamiliar token, such as a nonword that does not have exemplars, imposes other constraints that may override typical neighborhood density effects on articulation.

However, these durational effects do not generalize uniformly and are more contested. Notably, Munson and Solomon (2004) found no significant effect of phonological neighborhood density on vowel duration in English speech, despite PND effects on vowel-space expansion. Gahl et al. (2012), however, report shorter durations for words from dense neighborhoods in

spontaneous speech. This discrepancy suggests that PND effects on duration may rely on factors beyond lexical access.

While much of the research on PND effects in speech production has concentrated on vowel articulation and durational properties, a growing body of evidence suggests that PND also influences consonantal realization, particularly in terms of segment duration and coarticulation. In Polish, fricatives in words with high PND exhibited increased fricative durations compared to those in words with low PND (Kaźmierski, 2019). As with vowels, this is interpreted as a strategy to enhance the perceptual distinctiveness of words that face a greater risk of confusion for the listener.

High neighborhood density has also been associated with greater coarticulation in French speech production (Scarborough, 2018). Scarborough interprets nasal coarticulation as a listener-oriented strategy, since coarticulatory cues enhance contrast with competitors in dense PND neighborhoods. Oral vowels in high PND words showed the expected vowel space expansion, while nasal vowels in high PND words were centralized. The contraction in vowel space for nasal vowels is attributed to strong nasalization, which causes vowels to be pronounced more centrally, and thereby exaggerates contrast with their oral counterparts (Scarborough, 2018). This finding suggests that PND-driven hyperarticulation and hypoarticulation, in terms of vowel space expansion and contraction, is likely language-specific.

Based on PND and WF effects in human speech, researchers have begun to investigate how AI-generated speech may exhibit similar effects. Song et al. (2025) used Amazon Polly's Text-to-Speech (TTS) model (Amazon Web Services, 2024) to test for PND- and WF-based acoustic effects in synthesized speech. Song et al. selected target words varying in both PND and WF and had Polly read them at normal and fast speaking rates, measuring vowel duration and F1/F2 formants. They found no significant effect of PND or WF on Polly's vowels. Speaking rate was the only factor that produced predictable durational changes. The results indicate that Polly's speech generation does not reflect the same lexical or phonological sensitivities as those of humans. Song et al. interpret these results as evidence that Polly lacks the cognitive processes that underlie human speech planning — a limitation they characterize as the absence of a "theory of mind" for speech production.

1.4. Word Frequency (WF)

Word frequency, or lexical frequency, is a measure of the rate at which a word occurs in a corpus and is used extensively in word processing and psycholinguistic research (Brysbaert & New, 2009). Among the varying metrics used, raw frequency is the total count of occurrences of a word in a corpus and, as a result, scales with corpus size.

Frequency per million normalizes for corpus size by dividing raw frequency counts by the total number of tokens in the corpus and multiplying by one million, allowing comparison across corpora. The Zipf score is a logarithmic frequency measure ranging from approximately 1 (low-frequency) to 7 (high-frequency) (van Heuven et al., 2014). The logarithmic scaling reflects the distributional property captured by Zipf's law. In a given corpus, the most frequent word occurs approximately twice as often as the second-most frequent, three times as often as the third-most frequent, and so forth, capturing a logarithmic frequency distribution of words.

1.4.1. Word Frequency Effects on Speech Perception

Higher word frequency is associated with faster response times and improved accuracy in both visual and auditory word recognition (see Brysbaert & New, 2009). This facilitation is typically attributed to the stronger lexical representations that frequent words accumulate through repeated exposure, which consequently demand less processing effort for their identification (Stephenson, 2004). This effect depends on whether the frequency is encountered in L1 or L2. In L1, high-frequency words are more deeply entrenched in the lexicon and less susceptible to interference from competing lexical representations. The reduced exposure to L2 words weakens this entrenchment, and the magnitude of the word frequency effect is correspondingly larger in L2 visual word recognition (Diependaele et al., 2013; Brysbaert et al., 2017).

In bilingual word recognition, lexical frequency influences processing in both languages simultaneously. Dijkstra et al. (1999) showed that for homographs and cognates in L1 and L2, visual lexical decision times depend on the word's frequency in both the target and non-target language. For bilinguals of Dutch and English, low-frequency words in one language but high-frequency words in the other produce cross-language interference. This pattern is consistent with models in which lexical representations from both languages are co-activated during recognition (Dijkstra et al., 1999).

The lexical status of a word — for example, if it is a content word (e.g., *dog*) or a function word (e.g., *to*) — affects how frequency influences its duration (Bell et al., 2009). For content words, higher frequency is associated with shorter durations, and this effect is independent of other factors such as predictability. For function words, the apparent frequency-duration relationship is largely absorbed by previous-word predictability and speech rate, leaving only a minimal residual effect of frequency itself.

In AI speech recognition, WF appears to affect accuracy in approximately the same manner as humans, but with some nuances. Song et al. (2025) found that Alexa recognized high-frequency words more accurately than low-frequency ones, paralleling the human pattern. However, the Alexa model and humans may have different WF metrics as training data exposure does not match human language exposure. For example, Alexa recognized the word *dune* with higher accuracy than expected, likely because its training data contained many references to the recent *Dune* films (Song et al., 2025). This study suggests that while WF effects in AI systems mirror the general direction of human effects, the underlying frequency distributions reflect training data rather than human language exposure.

1.4.2. Word Frequency Effects on Speech Production

In addition to facilitating word recognition, high WF also affects speech production by increasing vowel duration and expanding vowel space (Umeda, 1975). Word frequency influences the acoustics of vowels through vowel space expansion and dispersion.

These production effects are explained by the Hypo- and Hyperarticulation (H&H) theory (Lindblom, 1990), which asserts that speakers continuously balance the demands of the listener for clarity with their own with articulatory economy. However, word frequency effects in speech production vary with lexical and phonological factors. PND and WF databases are derived from large language corpora that provide lexical and frequency metrics. An important methodological consideration in constructing such databases is the modality of the data in the corpora. In terms of WF, spoken and written language differ systematically, sometimes to a high degree, in their word distribution and frequency profiles. Spoken corpora, particularly spontaneous speech, contain more conversational vocabulary and phonological reductions (e.g., grammatical contractions and shortened high-frequency function words), whereas written corpora tend to overrepresent more formal literary words (Ernestus & Warner, 2011). Brysbaert & New (2009) found that WF data

based on television and film corpora better represent actual language use than written text corpora, at least with regard to relatively short words (one or two syllables) in English. However, relying on a single modality can skew word frequency estimates. Additionally, since every speaker in a language encounters words in different contexts and at varying frequencies, and individual speakers differ in their lexicons, this variability across speakers shapes individual lexical representations. In this sense, the most accurate metric of WF is speaker specific. Without knowing the comprehensive linguistic input of a single individual — an arguably impossible task — an exact WF metric is unknowable. Although canonical word frequencies are, to some extent, estimates, the vast and growing digitization of language in the last several decades has allowed researchers to calculate more representative figures.

Corpora play an immense role in how WF is studied, and the available corpora shapes what languages and phenomena get studied. There is a predominance of English in corpora — and in linguistics, experimental psychology, and related fields more broadly. The majority of word stimuli in experimental psychological research are in English (Balota et al., 2007), which raises questions about the generalizability of findings to other, less-studied languages, such as Polish. English-language corpora include the Corpus of Contemporary American English (COCA; Davies, 2008), the British National Corpus (BNC; BNC Consortium, 2007), and the SUBTLEX family of subtitle-based frequency corpora (SUBTLEX-US: Brysbaert & New, 2009; SUBTLEX-UK: van Heuven et al., 2014). Drawing on SUBTLEX frequency data, CLEARPOND (Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities; Marian et al., 2012) was developed to provide orthographic and phonological neighborhood statistics for multiple languages, but does not yet include Polish. Existing corpora for Polish include the SUBTLEX corpus (SUBTLEX-PL; Mandera et al., 2014), which provides subtitle-based frequency estimates derived from millions of words in Polish film and television subtitles. Yet, neither SUBTLEX-PL nor the National Corpus of Polish (NKJP; Przepiórkowski, 2012) includes PND data, which may partially explain the paucity of PND research in Polish.

1.5. Mixed Effects of Word Frequency and Phonological Neighborhood Density

The combined effects of WF and PND offer a more complete understanding of their impact on both perception and production. Vitevitch and Luce (1998) explored the distinct effects of phonotactic probability (which tends to facilitate word recognition) and neighborhood density (which tends to impede it), proposing a theoretical distinction that attributes these effects to sublexical and lexical representations, respectively. The combination of both PND and WF creates a combined effect (Wright 2004; Munson & Solomon, 2004; Scarborough & Zellou, 2013). This interaction is described by distinguishing "easy" words (characterized by high word frequency and sparse phonological neighborhood density) from "hard" words (defined by low word frequency and dense phonological neighborhoods). Words with mixed characteristics — high word frequency with dense neighborhoods or low word frequency with sparse neighborhoods — tend to display more balanced results, with the effects canceling each other out. Figure 1.5 summarizes the combined predicted acoustic effects of phonological neighborhood density and word frequency.

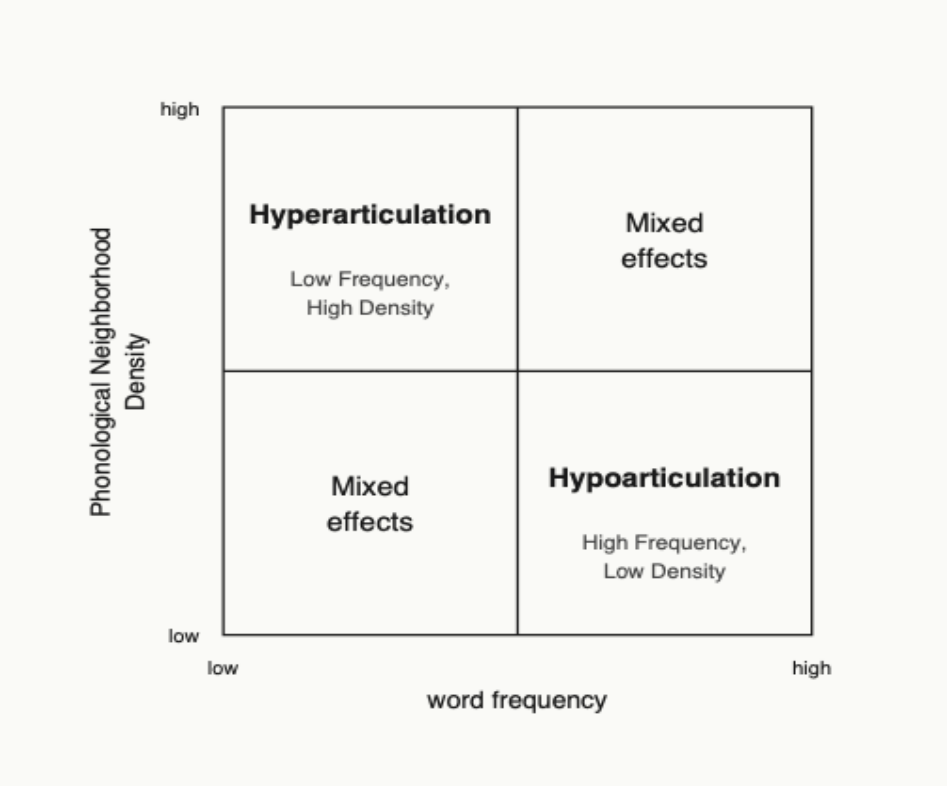

Figure 1.5. Combined Effects of Phonological Neighborhood Density and Word Frequency (author-created)

.

Across numerous studies, listeners perceive easy words with greater accuracy and faster reaction times than hard words (Bradlow & Pisoni, 1999; Wright, 2004). This pattern holds true for both native and non-native listeners, although non-native listeners often exhibit a more pronounced difference in recognition accuracy between these two categories, suggesting a heightened sensitivity to these lexical variables under reduced proficiency (Bradlow & Pisoni, 1999).

The interaction between lexical frequency and speech perception is paradoxical. While expanded vowel spaces are generally linked to increased speech intelligibility (Bradlow et al., 1996), studies have consistently shown that these hard words are still harder to perceive than easy words (Luce & Pisoni, 1998). This suggests that the perceptual challenge posed by low word frequency and dense phonological neighborhoods may be more potent than the compensatory articulatory efforts from the speaker (Vitevitch & Luce, 1998). Therefore, inhibitory and facilitatory effects are shaped by both acoustic-phonetic and lexical factors, making it difficult to isolate the contribution of lexical frequency from that of phonological context.

Wright (2004) and Umeda (1975) have demonstrated that hard words — characterized by dense phonological neighborhoods and low-frequency words — are produced with more expanded vowel spaces, indicative of greater vowel dispersion or hyperarticulation, when compared to easy words that are high-frequency and in sparse phonological neighborhoods. However, there are other methods to calculate vowel space expansion in addition to measuring Euclidean distance from the centroid. Stephenson (2004) calculated inter- and intra-vowel distances as other methods by which to investigate vowel dispersion. Inter-vowel distance is defined as the mean Euclidean distance between each vowel token and every token outside its vowel category. When averaged across all tokens, this yields a single value for hard words and one for easy words, capturing the full pattern of acoustic separation across the vowel system. This method is more sensitive to system-wide contrast and overlap than centroid-based approaches. Hard words showed greater inter-vowel distance than easy words, further supporting the idea that hard words are more dispersed in vowel space, particularly at the system-wide level.

Stephenson (2004) then measured intra-vowel distance — the mean Euclidean distance between all tokens within the same vowel category — to quantify in-category variability. Each token was compared to every other token within its category, and these distances were averaged to yield a single measure of within-category dispersion. Unlike inter-vowel distance, which reflects how far apart vowel categories are, intra-vowel distance captures how tightly tokens cluster within a category. Easy and hard words showed no significant difference in this measure, suggesting that while hard words expand the vowel space overall, their tokens are not more variable within individual categories.

1.6. Linguistic Corpora

Phonological neighborhood density (PND) and word frequency (WF) databases are derived from large language corpora that provide lexical and frequency metrics. An important methodological consideration in constructing such databases is the data modality in the corpora. In terms of WF, spoken and written language differ systematically, sometimes to a high degree, in their word distribution and frequency profiles. Spoken corpora, particularly spontaneous speech, often capture more conversational vocabulary and phonological reductions (e.g., grammatical contractions and shortened high-frequency function words), whereas written corpora tend to

overrepresent more formal literary words (Ernestus & Warner, 2011). Brysbaert & New (2009) found that WF data based on television and film corpora better represent actual language use than written text corpora, at least with regard to relatively one- or two-syllable words in English. However, relying on any one modality can skew WF estimates. Since every speaker in a language encounters words in differing contexts and frequencies, and individual speakers differ in their lexicons, this variability across speakers shapes individual lexical representations. In this sense, the most accurate metric of WF is speaker specific. Therefore, without knowing the comprehensive linguistic input of a single individual — an arguably impossible task — an ex act WF metric is unknowable. Although canonical word frequencies are, to some extent, estimates, the vast and growing digitization of language in the last several decades has allowed researchers to calculate more representative figures.

To generate PND metrics, words in a corpus are first converted from orthographical forms to phonemic transcriptions, typically by means of pronunciation dictionaries or through forced alignment techniques applied to speech data. Each word's phonological form is then compared to others in the lexicon to determine the number of neighbors and the density of the neighborhood, as discussed above. This conversion differs across languages, as some (e.g., English) have complex or inconsistent orthographies that require more robust orthographic-to-phonemic conversions, while others (e.g., Polish) have consistent orthographies that streamline the conversion.

While there is a Polish SUBTLEX corpus, which provides subtitle-based frequency metrics derived from millions of words in Polish film and television subtitles, it lacks PND metrics. However, it is not used as frequently as the National Corpus of Polish (NKJP) (Przepiórkowski, 2012). While the NKJP contains over 1.5 billion words, neither it nor any of the other historical Polish corpora contains PND data, which may partially explain the lack of PND research in Polish. Additionally, since the NKJP is based primarily on written sources, it likely overrepresents formal vocabulary.

Prior to Alzahrani's (2025) Jiwar database, no dedicated database of Polish PND was publicly available, and neighborhood densities had to be computed on an ad-hoc basis. This is an attempt to solve this issue for underrepresented languages, including Polish, by creating an open-source database of neighborhood density and frequency metrics for 40 languages. The scope of Jiwar database allows for more consistent comparisons of these metrics across languages. Additionally, that data are drawn from TV/movie subtitles rather than written sources, which tend

to reflect more accurate WF metrics overall (Brysbaert & New, 2009). This thesis uses the Jiwar database to explore the effects of PND and WF in Polish, with the aim of selecting the most appropriate metrics for creating stimuli in the experiments.

Chapter 2: Modeling Speech Using Artificial Intelligence

2.1. Artificial Intelligence and Neural Networks

This thesis addresses both linguistics and artificial intelligence, aiming to offer novel contributions to both disciplines. The second chapter situates contemporary speech synthesis and speech recognition systems within the broader field of artificial intelligence (AI) and language modeling. It addresses the theoretical and historical grounding of the modern effort to develop ever more sophisticated and efficient models for speech recognition and synthesis. The first sections introduce neural network architectures and their core principles. Subsequent sections describe the specific approaches to speech synthesis and speech recognition that provide context for the AI models examined in Chapters 4 and 5: Amazon Polly, ElevenLabs, and Meta's wav2vec 2.0.

The computational models with the most direct biological inspiration are artificial neural networks, which are based on simplified mathematical abstractions of biological neural systems (Rosenblatt, 1958). Neural networks offer a flexible architecture for approximating complex, nonlinear functions. When neural networks contain many layers, enabling the hierarchical extraction of features from raw data, they are referred to as deep learning models (LeCun et al., 2015). Deep learning has become synonymous with much of modern machine learning (ML) and is present in the AI applications that first crossed over into popular use.

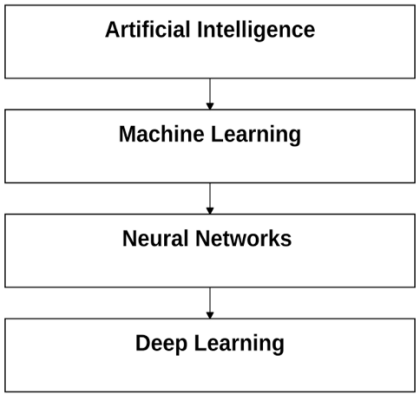

Figure 2.1. Taxonomy of AI terminology (author-created).

Several terms recur throughout this thesis and require disambiguation. *Deep Learning* refers to any neural network with multiple layers trained to extract hierarchical features from data. *End-to-end* is a modeling approach that maps raw input directly to the final output, as opposed to more traditional modular/multi-stage methods. *Attention-based* refers to any architecture that uses an attention mechanism, in which the model "attends" to certain elements more than others by dynamically weighing the relationships between elements of the input sequence. Though attention mechanisms predate the transformer architecture, and were used in earlier recurrent neural network models, they are often used almost synonymously with *transformers*: a type of neural architecture introduced by Vaswani et al. (2017) that extended the notion of attention by implementing self-attention, positional encoding, and other mechanisms. While all transformers are attention-based, not all attention-based models are transformers. Similarly, while most contemporary end-to-end systems are transformer-based, the terms end-to-end and transformer refer to distinct properties. All four terms apply to the speech systems studied in this thesis (Polly, ElevenLabs, and wav2vec 2.0). Where finer distinctions matter, the more specific term is used.

The terms "artificial intelligence" and "machine learning" emerged in the 1950s. Both initially lacked clear and universally accepted definitions, though AI was broadly understood as the simulation of human intelligence, and machine learning as the study of algorithms. After early

conceptual breakthroughs, progress stalled for several decades, as computational power and data availability constrained practical advances. These limitations contributed to "AI winters" of declining interest and confidence in the discipline (Toosi et al., 2021). A resurgence in the late twentieth and early twenty-first centuries, driven by the compilation of large datasets, advances in statistical methods, and increased computing power, transformed machine learning into a central pillar of modern AI research and applications (Toosi et al., 2021; Dean, 2022).

Arguably, the most decisive shift in AI occurred in 2017 with the publication of *Attention Is All You Need*, which introduced the transformer architecture (Vaswani et al., 2017). Transformers largely replaced previous neural network approaches such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) models. It leveraged attention and transformer mechanisms to enable parallel processing of entire sequences, leading to efficient scalability on GPUs and better capture of long-range dependencies within data. Initially applied to English/German translation by Vaswani et al. (2017), the attention-based transformer architecture quickly demonstrated broader applicability, revolutionizing natural language processing and becoming the foundation for subsequent advances in large-scale generative models. It was with this technology that generative AI entered into widespread public usage through LLMs, including OpenAI's ChatGPT, Anthropic's Claude, Google's Gemini, and Meta's Llama.

These developments are the source of broader conceptual debates. Since the explosion in interest brought on by the advent of LLMs, the theoretical notion of Artificial General Intelligence (AGI) has gained prominence. Definitions and thresholds for AGI vary, but the term broadly refers to a hypothetical system capable of human-level reasoning and adaptability across diverse tasks. Despite its visibility in both industry and popular discourse, AGI remains a contested and ill-defined concept, with many researchers regarding it as aspirational or speculative rather than imminent (Mitchell, 2021).

The evolution of AI has progressed from symbolic to connectionist, and from very basic neural networks to compute-intensive deep learning. While AI remains the umbrella term that is familiar to most, it is actually deep learning — and more specifically, the attention-based transformer networks on which it runs — that is currently driving the field's most significant

advances. The modeling aspects of this project leverage the architecture of neural networks to explore linguistic similarities and differences between human cognition and AI.

2.1.1. Neural Networks and Deep Learning Fundamentals

At its most essential level, deep learning aligns multi-layered neural networks to patterns in data through the training process. While artificial neural networks are inspired by the neural systems of organisms, they are mathematical rather than biological constructs. The most basic building block of a neural network, the neuron, holds a numerical value known as an activation, which is a numerical representation of how much that neuron is firing. Neurons are arranged in several types of layers (input, hidden, and output), where each neuron in one layer is connected to all neurons in the following layer (Schmidhuber, 2015). The strength of the connections between neurons is governed by parameters known as weights and biases. Neuronal activation is computed as the weighted sum of the preceding layer's activations, adjusted by a bias term, and then passed through a nonlinear activation function (Schmidhuber, 2015).

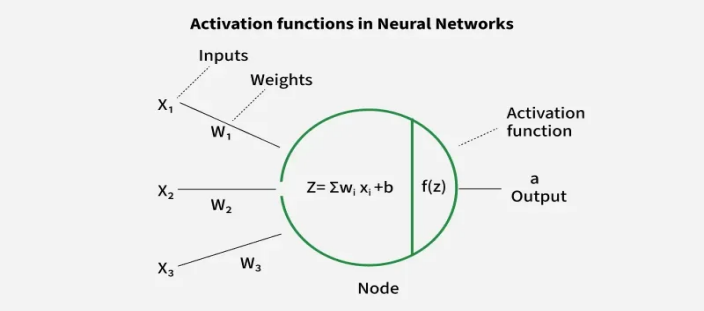

Figure 2.2. Activation functions in neural networks (author-created).

With the exception of the input layer, each node in the network takes the weighted inputs from the neurons of the preceding layer, processes them through a linear transformation with a summation, and applies a nonlinear activation function to create the output. In Figure 2.2 above, for a node receiving an input vector $x = [x_1, x_2, \ldots, x_n]$ and weight vector $w = [w_1, w_2, \ldots, w_n]$, with the addition of a bias term *b*, the neuronal output is computed as $y = f(w^T x + b)$ where f() is an activation function such as a sigmoid (σ), hyperbolic tangent (*tanh*), or rectified linear unit (ReLU).

While larger networks are much more complex, the archetypal neural network can be represented as shown in Figure 2.3.

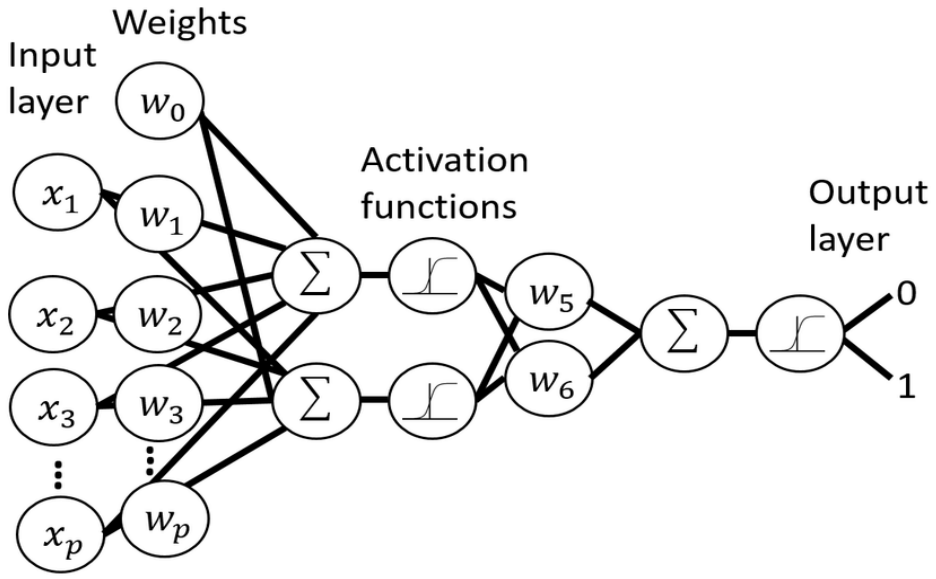

Figure 2.3. Diagram of a simple neural network (author-created).

Activation functions add nonlinearity to the network, which enables it to learn and represent the complex, nonlinear relationships that occur in real-world data, such as acoustic speech signals. Early networks often employed the sigmoid activation, but at present, modern deep learning networks use other activation functions to avoid vanishing gradient problems that occurs at the extremes of the sigmoid's input range. The choice of activation function determines how signals propagate through the network and how effectively it can learn complex mappings. There are numerous activation functions, but the following sections describe those most relevant to this thesis.

The sigmoid function transforms input values into the range of 0 to 1 using: $\sigma(x) = \frac{1}{1 + e^{\{-x\}}}$where x is the input value. This produces a long S-shaped curve, and has been used in traditional binary classification models in which outputs represent probabilities. A significant drawback is its tendency to exhibit vanishing gradients as output values for large values of x. The sigmoid function maps any input to a value between 0 and 1, as shown in Figure 2.4.

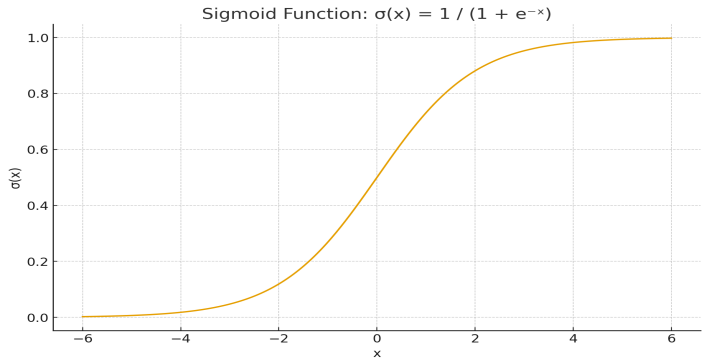

Figure 2.4. The sigmoid function (author-created).

The hyperbolic tangent (*tanh*) function maps inputs to values from −1 to 1 and is expressed as: $tanh\ (x)\ =\ \frac{e\^{}\{x\} - e^{(-x)}}{e\^{}\{x\} + e^{(-x)}}$. While the sigmoid function maps any value to another value between 0 and 1, *tanh* retains the positive/negative sign of the input by mapping all negative inputs from -1 up to 0, and all positive inputs from 0 up to 1. However, *tanh* still suffers from vanishing gradients when inputs are extreme. Figure 2.5 shows the *tanh* function, which produces outputs in the range [−1, 1].

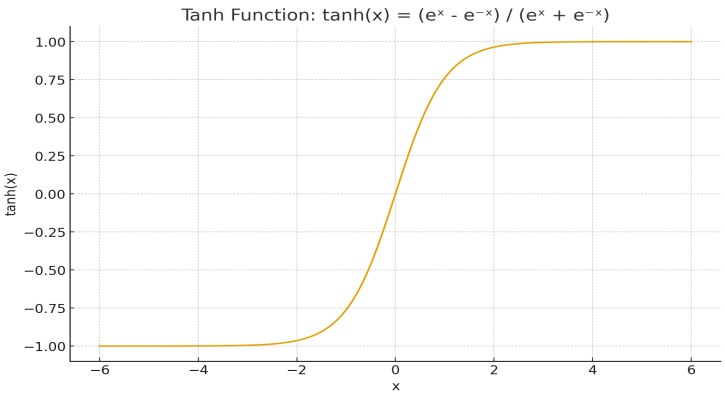

Figure 2.5. The *tanh* function (author-created).

The rectified linear unit (ReLU) maps all negative inputs to zero and increases linearly for positive inputs: $f(x) = \max(0, x)$. ReLU has now become the standard activation function in deep convolutional and transformer networks, as it can mitigate vanishing and exploding gradients. This

is in spite of the other issues associated with ReLU, including inactive neurons in cases where many outputs remain at 0, which is illustrated in Figure 2.6.

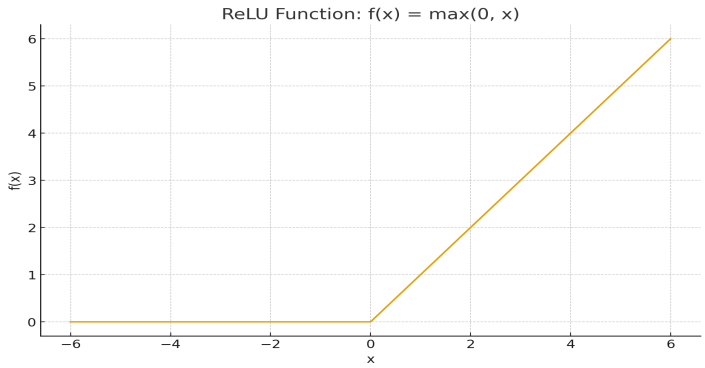

Figure 2.6. The ReLU function (author-created).

The function that is most commonly used in deep learning neural networks to convert large vectors of unbounded numbers into a normalized probability distribution (i.e., 0 to 1) is the softmax function, expressed as $softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$, as depicted in Figure 2.7. The softmax function is widely used in the output layer of attention-based transformer models.

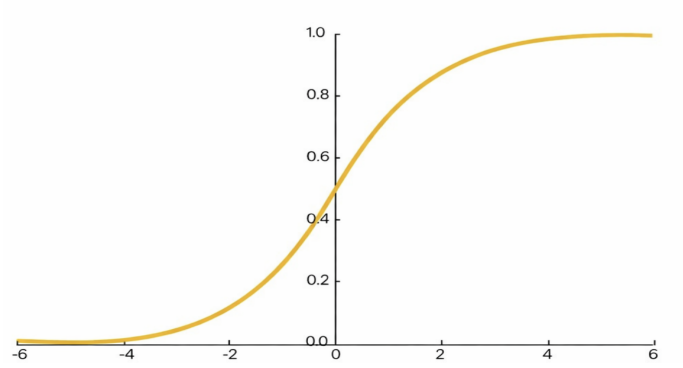

Figure 2.7. The Softmax function (author-created).

2.2. Conceptual Computational Framework

The field of artificial intelligence is a math-heavy domain due to its statistical nature. Interacting with these architectures at a deeper level requires knowledge of linear algebra, calculus, and statistics, as well as insight into training theory and methodology.

2.2.1. Linear Algebra and Vector Spaces

Linear algebra provides the structural and computational framework upon which modern AI systems are built. Data are represented in the form of vectors and matrices, which allow for efficient manipulation and transformation (Jurafsky & Martin, 2023). These vectors are mathematical representations of positions in geometric space (theoretical space with more dimensions than the physical three-dimensional world upon which the concept is based). Operations such as linear transformations, projections, and tensor products provide methods to encode, compress, and extract features, as well as to interpret complex relationships within data. These mathematical tools underpin core algorithms for dimensionality reduction, optimization, and neural network computation, and they enable the scalable implementation of models that learn from vast datasets. Linear algebra provides the ability to conceptualize data geometrically and display data relationships as the manipulation of points and vectors in high-dimensional space.

In speech modeling, linear algebra is foundational to both recognition and synthesis. Acoustic signals are converted into feature vectors, which are then arranged into matrices or tensors for processing by neural networks. Embeddings of phonemes, words, and acoustic segments are defined within high-dimensional vector space, where geometric properties such as distance and angle correspond to linguistic similarity. When speech data are modeled with AI, the principles of linear algebra determine how they are represented and transformed.

2.2.2. Vectors and Data Representation

Among the many uses of vectors in AI and natural language processing, one of the most prominent is in Large Language Models (LLMs). These models learn vector representations of linguistic units, known as embeddings (Devlin et al., 2019). In deep learning, embeddings are vectors that encode items as points in a high-dimensional representation space defined by latent features. More generally, vectors are used throughout neural networks to represent inputs, intermediate activations, and model parameters, including weights. A vector is depicted as an ordered list of numerical values in Figure 2.8.

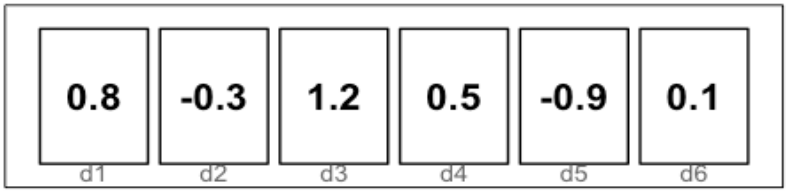

Figure 2.8. A basic example of a vector (author-created).

Vector mathematics allows multiple operations and transformations to be performed efficiently, often in parallel. These include vector addition, subtraction, scalar multiplication, and dot products. A vector representing a word consists of values across multiple dimensions, each corresponding to a learned latent feature in a high-dimensional representation space. The geometric interpretation of these vectors as embeddings allows for direct comparison between words. For example, the Euclidean distance or cosine similarity between embeddings can be used to quantify similarity across dimensions. This enables LLMs and related models to represent and process linguistic structure in a mathematically tractable way. Embeddings for words such as *uncle* and *aunt* are typically closer to each other than to unrelated words like *car*. Relationships between words are often captured as vector offsets: the difference between *uncle* and *aunt* is similar to the difference between *man* and *woman,* and *king* − *man* + *woman* is close to *queen* (Nikolov et al., 2013).

Measuring the similarity between two words and applying the result in another mathematical operation to obtain a similarity score is a key operation for a network to compute attention in transformer models such as LLMs. Simple vector subtraction yields further information about similarity and feature representation in a high-dimensional space. Impressionistically, the semantic difference between gendered words (e.g., *man*/*woman, boy/girl, aunt/uncle)* leads to the assumption that gender is a feature or dimension that separates these words in a feature space (Mikolov et al., 2013).

Representing words, or more precisely tokens, as embeddings is only one of many uses of vectors in AI. Vectors are used to represent features, activations, and model parameters (e.g., weights) within neural networks. While vectors represent individual points in high-dimensional space, matrices and tensors organize values into structured arrays that support efficient computation. A matrix can be understood as a collection of equally sized vectors arranged as rows or columns, where each vector encodes values across a shared set of features. The output of a

neural network layer is typically computed through matrix multiplication followed by nonlinear transformations (activation functions), applied repeatedly across layers. Organizing computations in terms of matrices and tensors is not only a mathematical convenience but also enables substantial computational gains through parallelization.

Hardware advances are a key aspect of this parallelization of linear algebra operations that has allowed AI to advance so rapidly. The hardware revolution both reflects the increasing transistor density described by Moore's Law (Moore, 1965), and the repurposing of Graphics Processing Units (GPU) from their original graphics role to training neural networks (LeCun et al. 2015). GPUs were originally developed as complementary processing units for Central Processing Units (CPU) to accelerate the matrix and vector transformations in computer graphics. To accomplish this, GPUs consist of many small, special-purpose parallel processing units, which perform linear algebra operations simultaneously and in parallel. Because neural networks rely on parallel vector and matrix operations, as do computer graphics, adapting GPUs for deep learning operations has led to substantial improvements in the speed of neural network training (LeCun et al., 2015). In contrast, CPUs, designed for general-purpose, serial processing, are considerably slower for neural network workloads than GPUs (Owens et al., 2007).

2.2.3. Learning in Neural Networks: Gradient Descent and Backpropagation

The primary mechanism of learning in a neural network involves adjusting the weights associated with each neuron so as to reduce the difference between its real outputs and the desired outputs. The weights are typically initialized to random values, and then adjusted incrementally as data propagates through the network. In gradient descent, the network learns to more correctly map from input data to target outputs by iteratively updating its parameters during training. When the network's output differs from the target output, a loss function quantifies this error, and the gradients of this loss with respect to the network's parameters are computed via backpropagation. These gradients are then used to update the weights in a direction that reduces the error. This process is often visualized as navigating a loss landscape, where the goal is to reach a minimum. Although this landscape is high-dimensional in practice, it is commonly represented as a two-dimensional surface for simplicity, with the optimal parameter configuration corresponding to the lowest point. With each training iteration, the parameters are updated in the direction of the negative gradient, moving incrementally toward this minimum. As shown in Figure 2.9.

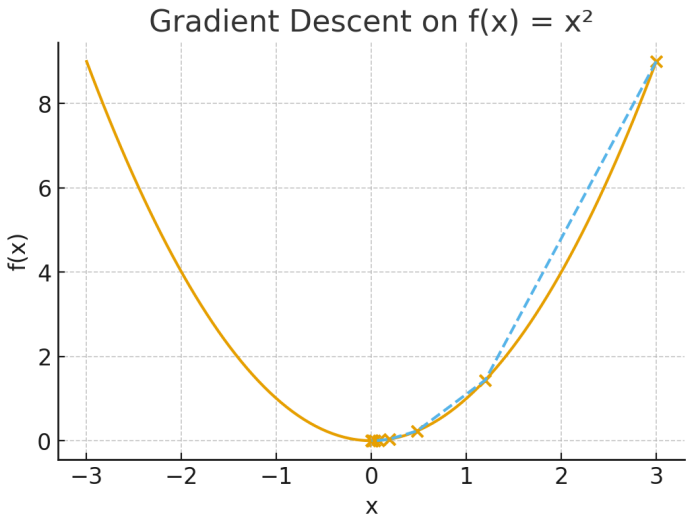

Figure 2.9. Gradient descent (author-created)

.

In this sense, gradient descent is effectively an algorithm that repeatedly adjusts the weights to move step-by-step downhill on the gradient curve to find the point where the function reaches its optimal value (Goodfellow et al., 2016). In the two-dimensional case of Figure 2.9, the problem is relatively straightforward. However, the difficulty of training deep neural networks stems from applying this optimization across many deep layers, resulting in an extremely complex, high-dimensional landscape. Conceptualizing this high-dimensional loss surface as a reduced 3-D landscape reveals its complications, as the loss landscape can contain valleys, plateaus, and cliffs. Even if the loss function is moving in the correct direction locally toward a downhill slope, it may end up in the wrong valley and not reach the optimal solution. A three-dimensional depiction of the loss surface and descent trajectory is provided in Figure 2.10.

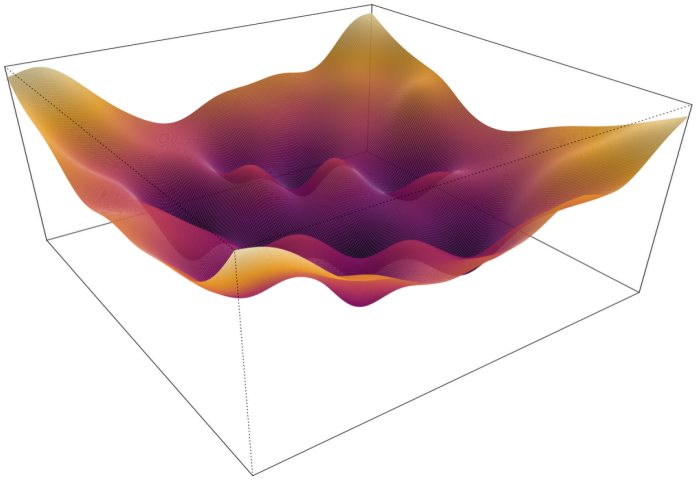

Figure 2.10. A graphical three-dimensional depiction of gradient descent (author-created).

In deep learning, the loss landscape can consist of thousands of dimensions, making the process of optimally adjusting the weights vastly more difficult than in the case of a two- or three-dimensional curve. The complexity involved in this optimization creates two types of gradient descent problems in deep learning — vanishing and exploding gradients. When gradients vanish, the slope of the loss function disappears as it is back-propagated through layers of the network until the gradient descent no longer moves along the landscape. In vanishing gradients, network learning slows or stops entirely, and the network does not learn the optimal weights to learn the desired output. In the case of exploding gradients, the slope of the loss function becomes so large and excessively steep that the weight updates always overshoot the valleys, never finding the optimal weights to learn the desired output. Both cases occur because, in deep learning, gradients are repeatedly multiplied through many layers of the network, and therefore, the repeated multiplication can cause the weight values to vanish toward zero or explode toward infinity, as illustrated in Figure 2.11.

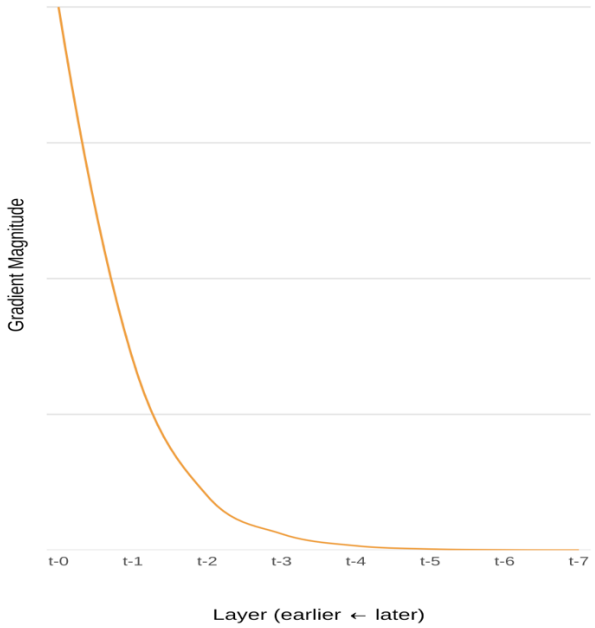

Figure 2.11. Exploding gradient where *t* is the time step

.

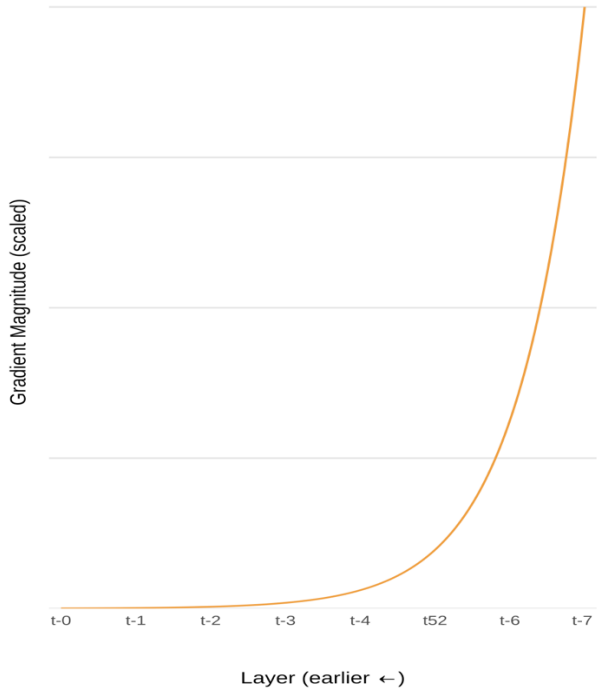

Figure 2.12. Vanishing gradient where *t* is the time step.

2.2.4. Embedding Spaces

Embedding vectors typically consist of hundreds or thousands of dimensions (e.g., OpenAI's text-embedding-3-large model produces vectors with 3,072 dimensions), where each dimension corresponds to a component of a high-dimensional representation. In the case of linguistic and semantic structure, this means that there is no way to interpret individual dimensions as corresponding to explicit symbolic features; instead, these structures are encoded across combinations of dimensions.

During training, embedding representations are learned and updated to capture patterns of usage based on the context and relationships between tokens in the training data. Because these embeddings are structured geometrically, relationships between tokens can be represented as relative positions or vector offsets in this space. As a result, geometric analysis of embedding spaces can reveal latent structure, though this structure is not, at present, directly interpretable. The interpretability of embedding spaces remains an active area of research, including work on phonetic relationships within such spaces.

The geometry of embedding spaces raises theoretical issues in psycholinguistics and cognitive science. Although such representations can simplify and clarify aspects of model behavior, a core research question concerns whether the structure of these spaces aligns with the representations underlying human cognition and perception. For example, there is ongoing research into the degree to which the encoded representations correspond to perceptual distances or articulatory dimensions in ways that meaningfully reflect the structure of vowel space. Even if sufficiently large models can capture linguistic regularities from large-scale text and acoustic data, an open question remains regarding the extent to which these embedding spaces align with human cognitive representations, if at all.

While deep learning models such as LLMs are not intended to be a precise simulation of neuronal processes in the brain, some research suggests that certain common geometric patterns exist between the artificial contextual embeddings of LLMs and brain embeddings. Goldstein et al. (2024) suggest that language areas in the brain may rely on a gradient embedding space to represent language, which is a departure from discrete, symbolic representations posited by traditional theories. In Goldstein et al., during listening tasks, this geometric alignment appears to be precise enough that the pattern of neuronal activity in the inferior frontal gyrus in a subject can be predicted using only its geometric relationship to other words within the embedding space.

Embeddings may also capture dimensions that align with human conceptual systems. Grand et al. (2018) propose that word embeddings can reveal nuanced, context-dependent human judgments about features like size, intelligence, and danger.

In speech recognition, embeddings are essential for representing both semantic and acoustic information. Phonemes or subword units can be embedded in high-dimensional spaces such that acoustically or articulatorily similar sounds are geometrically proximate. For example, the English vowels /i/ (as in beet) and /ɪ/ (as in bit) are perceived to be close to one another in vowel space because they share similar formant structures. By encoding these vowels as nearby points in a vector space, models can account for perception and acoustic variability, allowing recognition systems to generalize across speakers and contexts. Similarly, each acoustic frame (for example, 20-millisecond slices of speech) can be mapped into embedding spaces where spectral features are compressed into representations suitable for phoneme and word prediction.

In speech synthesis, embeddings process linguistic input into acoustic output by mapping words, subwords, and phonemes into embedding vectors that capture both categorical identity and contextual nuance (Ning et al., 2019). Prosodic or speaker-specific embeddings can also be incorporated, enabling systems to control for intonation, rhythm, or voice characteristics. By learning embeddings that encode these features, synthesis systems can generate natural speech. The flexibility of embedding-based representations makes it possible to extend synthesis models to new languages, voices, or speaking styles with minimal additional training (Ning et al., 2019).

While the high-dimensional embedding space allows for the discovery of novel relationships between tokens, several issues prevent the direct visualization of embeddings. By nature, humans have difficulty visualizing data in more than three dimensions. In order to do so, embeddings that contain thousands of dimensions must be reduced to just three, which results in an oversimplification of features and relationships. One common method of dimensionality reduction is Principal Component Analysis (PCA), which finds orthogonal directions — principal components — that capture the greatest variance in the data (Jolliffe, 2002). By projecting high-dimensional embeddings onto the top components, such as the first two or three (e.g., in a 2D or 3D visualization), PCA enables lower-dimensional visualizations.

2.3. Statistical and Neural Approaches to Modeling Language

Though many of these are niche or out of common usage, the following approaches informed the development of speech technology in ways that are critical to a full understanding of the field. The evolution of these approaches has direct bearing on the computational approach taken herein, particularly in terms of the inspection of internal activations.

2.3.1. Hidden Markov Models

Though Hidden Markov Models (HMM) are not neural networks, they dominated AI for decades during a period in which there was insufficient computational power and methodology for neural networks to be practical. An HMM is a probabilistic machine learning model designed for modeling sequential data. HMMs are based on the Markov property (Markov, 1906; see Jurafsky & Martin, 2023), which states that the probability of transitioning to the next state depends only on the current state. HMMs provide a powerful statistical approach for modeling sequential data, where observations are generated from an underlying sequence of states.

One of the most successful early applications of HMMs was speech recognition. In this framework, an audio signal is treated as a sequence of observable inputs, often represented as time-ordered observation vectors (e.g., 10-ms acoustic windows). The HMM then infers the most likely sequence of hidden states, which are not directly observable, but correspond to underlying linguistic units. By selecting the sequence of states with the highest probability, an HMM produces the most probable transcription of the spoken input (Jurafsky & Martin, 2023).

A key aspect of HMMs, which is both an advantage and a disadvantage, is that their statistical dependencies are local (Gales & Young, 2008). Each hidden state depends only on the previous state, and each observation depends only on the current hidden state. In other words, a given state at time *t* is always dependent on the state at time *t-1*, which restricts the model's ability to capture long-range dependencies in sequential speech data. For instance, in sentences with nested noun phrases (e.g., "People who love to read books prefer to go to bookstores"), an HMM may misinterpret structural relationships because it considers only the immediately preceding state, obscuring the true subject–verb dependency. This deficiency is a primary disadvantage in comparison to neural networks, whereas LLMs in particular can capture very long-range dependencies. However, the probabilistic inference methods of HMMs are more transparent than

neural networks, which are typically viewed as “black-box” models. HMMs also use fewer computational resources than neural networks, which proved a practical advantage in the late 20th century when computing power and memory were limited. This advantage was progressively eroded as deep learning networks were shown to outperform GMM-HMMs on a variety of acoustic modeling benchmarks (Hinton et al., 2012).

2.3.2. Recurrent Neural Networks

RNNs are highly flexible and capable of learning complex nonlinear data. Unlike HMMs, RNNs process sequential data while maintaining an internal representation of the information they have already encountered. The evolution of RNNs from traditional neural networks (e.g., feedforward networks or multilayer perceptrons) was notable for this innovation of memory for prior states. Recurrent loops in RNN architecture allow previous inputs to influence current outputs, creating long-distance interaction in the data. RNNs were also the first neural networks to be classified as deep learning, as they use multiple “deep” layers in their networks.

The ability of RNNs to model long-range dependencies came with significant costs. Training RNNs requires backpropagation, in which gradients are propagated backward across many time steps to update network weights. While this enables learning over longer sequences, the further from the current time step the model goes, the greater the risk of vanishing and exploding gradients. Both vanishing and exploding gradients prevent RNNs from effectively learning long-distance dependencies, which was the limitation that RNNs were designed to overcome. Additionally, because RNNs process tokens sequentially — one time step at a time — their computation scales linearly with sequence length and cannot be parallelized across time steps. This sequential structure requires the model to propagate information through recurrent connections over many steps.

RNNs were first developed in the 1980s (Rumelhart et al., 1986), with significant architectural advances in the 1990s (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997). However, early RNN architectures posed significant training challenges, and were often used in conjunction with HMMs rather than replacing them (Mehrish et al., 2023).

2.3.3. Long Short-Term Memory Networks

To solve the vanishing/exploding gradient problem, Hochreiter & Schmidhuber (1997) introduced a novel version of the RNN — Long Short-Term Memory (LSTM) network. The LSTM addressed the gradient problem by implementing several gates within each memory cell in the network. These gates (input, forget, and output gates) regulate input and output for each cell, reducing the gradient problem (Hochreiter & Schmidhuber, 1997; Gers et al., 2000). Unlike nodes or neurons in other types of neural networks, an LSTM cell maintains a hidden state across time steps through a set of learnable gates, allowing the network to selectively preserve relevant information over long sequences.

The forget gate is responsible for deciding what information should be retained and discarded from the previous state. This gate processes the previous hidden state and the current input using a sigmoid activation function (σ), which creates an output vector of values between 0 and 1. This output vector is then multiplied by the old cell state. If the value is closer to zero, the gate tells the node to "forget" the data, while a value closer to one tells the node to "keep" the data.

The input gate determines whether new information should be stored in the cell state. A sigmoid function selects what to keep, and a *tanh* function transforms the input and previous hidden state into candidate memory values. These filtered candidates are then added to the portion of the old cell state preserved by the forget gate, forming an updated cell state that combines past information with newly selected input. The outputs of both gates form the updated cell state C_t.

The output gate generates the next hidden state, which is passed to the next layer, where the process repeats, with the combined current input and previous hidden state passed through a sigmoid and a *tanh* function. These two results are multiplied to produce the new hidden state (h_t). This controlled filtering of the cell state is what allows LSTMs to preserve useful information and successfully capture long-term dependencies.

Despite their ability to control vanishing and exploding gradients better than earlier RNNs, LSTMs only mitigate the issue, not eliminate it. The additional gates and internal mechanisms also make LSTMs more computationally expensive to train and run. This extra complexity reduces interpretability, making LSTMs more opaque than previous architectures.

2.3.4. Attention-Based Transformer Networks

The introduction of the transformer by Vaswani et al. (2017) marked a watershed moment in AI. Transformer-based models replace recurrence with self-attention, processing all sequence positions in parallel. This eliminates the sequential bottleneck of RNNs, which substantially improves training efficiency and enables effective modeling of long-range dependencies. Self-attention also allows direct interactions between distant tokens, mitigating the gradient-propagation problems of RNNs. In combination with advances in hardware and optimization, this enabled models to scale to much larger datasets and begin generating highly fluent and coherent text.

The key to these advancements lies in the mechanism of self-attention and high-dimensional embedding space. Instead of processing tokens sequentially, transformer architectures compute relationships between all tokens in parallel using self-attention. The input is first tokenized, where a token may represent a word, subword unit, or segment of acoustic input. Each token is then mapped to a high-dimensional embedding, forming a learned representation derived from large training corpora. These representations consist of distributed, latent components rather than explicitly interpretable features. Through self-attention, the model computes weighted combinations of token representations, allowing it to capture contextual relationships across the entire sequence. This enables the model to represent each token in a way that reflects its context and its relationships to other tokens in the input.

These contextualized representations are iteratively updated across layers, allowing the model to distinguish between multiple senses of the same word — for example, *table* as a piece of furniture versus *table* as a structured set of data. Although this example focuses on text, similar context-dependent mechanisms can be applied to other modalities, such as speech or acoustic signals, within appropriately designed architectures. The contextualized representations from the final transformer layer in the output are mapped to logits. The model generates the softmax distribution for the next possible tokens at each step. Rather than always selecting the most probable token (which produces deterministic, repetitive output), models use sampling strategies that introduce controlled variability.

While the exact specifications of the latest commercial LLMs are not always publicly disclosed, well-documented models provide a sense of their scale. GPT-3, for example, uses 96 layers with 12,288 hidden units per layer, a context window of 2,048 tokens, and 175 billion

parameters (Brown et al., 2020). Model scale and capacity have continued to expand, driven by advances in hardware and training methods.

2.4. Speech Synthesis

Speech synthesis is the computational process of converting text into natural-sounding speech. This typically involves mapping orthographic input to phonemic representations, which are then realized as an acoustic signal. While lower-fidelity output may be acceptable for some applications, the ultimate goal is to produce natural, intelligible, and appropriately expressive human speech (Taylor, 2009).

Speech synthesis methods vary depending on language and application domain, but traditional speech synthesis systems are typically described as a three-stage process: text analysis, acoustic modeling, and waveform generation. In the first stage, the input text is normalized and transformed into a detailed linguistic or phonological representation that resolves ambiguities in pronunciation, prosody, and other implicit information, enabling accurate speech generation (Zen et al., 2009). For example, the English word *record* differs in stress depending on whether it is used as a noun or a verb, requiring lexical and syntactic disambiguation rather than grapheme-to-phoneme conversion alone. This stage is language-specific, as languages differ in their orthographic systems (e.g., alphabetic, logographic, syllabic) and phonological structure (Benesty et al., 2008). At the end of the first stage, the model has produced a fully specified linguistic representation — a roadmap for how to articulate the text, similar in function to IPA.

In the second stage, acoustic modeling uses the output of the first stage to predict the acoustic features required for speech generation. This stage maps the linguistic representation — containing prosodic, allophonic, and segmental information — onto a detailed acoustic representation suitable for synthesis. In traditional parametric systems, this process involves estimating features such as fundamental frequency (F0), amplitude, spectral characteristics, temporal dynamics, and coarticulatory effects (Tokuda et al., 2000).

In the third stage — waveform generation (or vocoding) — these acoustic features are used to generate a speech waveform in the time domain (Kawahara et al., 1999; Morise et al., 2016). In modern synthesis systems, elements of randomness are introduced into the speech signal for

variation in order to reduce the artificial/mechanical presentation of the synthesized speech. Figure 2.13 depicts the canonical three-stage pipeline on which many text-to-speech systems are based.

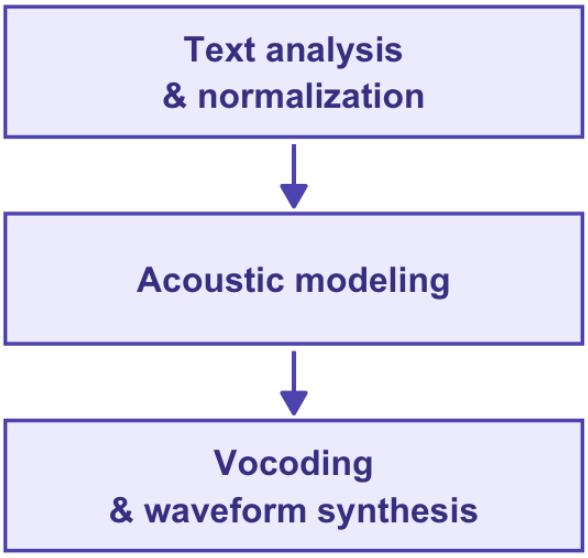

Figure 2.13. The canonical three-stage speech synthesis pipeline (author-created).

2.4.1. Articulatory Methods

Articulatory synthesis emerged in early speech synthesis research, in conjunction with advances in technologies such as X-ray and ultrasound. Based on modeling the vocal tract to approximate the physical processes that generate human speech, this approach simulates articulator movement, voicing, resonance, and aerodynamic properties rather than the resulting acoustic signal. The vocal folds are modeled as a vibrating source characterized by a fundamental frequency (F0), while turbulent noise sources represent the aerodynamic effects in the oral and nasal cavities necessary to produce sounds such as fricatives, allowing the model to reflect natural phonation and speaker-specific vocal characteristics (Klatt & Klatt, 1990). The primary drawback of this method is the difficulty of faithfully reproducing a synthetic version of a complex biological system.

2.4.2. Formant-Based Methods

Formant synthesis approaches, typically implemented as rule-based systems, differ from articulatory synthesis in that they model the acoustic output of speech rather than the underlying

physical processes of speech production. Instead of simulating the vocal tract and its articulatory gestures, formant synthesis represents speech using a source–filter method, in which the spectral envelope (i.e., formant frequencies and bandwidths) is generated through parameterized filtering.

In this approach, phonetic symbols are mapped to sets of acoustic parameters that define the behavior of the source and filter components, allowing the system to generate intelligible speech without explicitly modeling articulatory dynamics. Due to their relatively low computational and memory requirements, formant synthesis methods were widely used in early speech synthesis systems. However, despite achieving high intelligibility, these systems often produced speech that sounded unnatural. The large number of parameters required, along with the difficulty of tuning them accurately, particularly for speaker-specific characteristics, limited their overall performance.

2.4.3. Concatenative Methods

In the 1980s, concatenative synthesis became the dominant approach, as research shifted away from articulatory modeling and rule-based methods toward the assembly of speech from pre-recorded segments. This approach generally produced more natural and higher-quality output than earlier systems. However, it required substantial amounts of recorded speech — often many hours per speaker — and imposed significant storage requirements, which were constrained at the time. Its development was supported by advances in computing, including increased processing power and storage capacity.

The size of the stored speech units varies by implementation, and is a central design factor in concatenative systems, particularly in unit selection synthesis. Units may range from smaller segments such as diphones or syllables to larger units such as words or phrases, depending on the application. Larger units can improve naturalness by preserving coarticulatory and prosodic continuity, but they require more extensive databases. Smaller units reduce storage demands but make it more difficult to achieve smooth transitions, as natural speech exhibits significant coarticulatory variation. When adjacent units are not properly matched, spectral discontinuities can produce audible artifacts, reducing naturalness and intelligibility.

To address the challenge of choosing an appropriate speech-segment size, the most common approach in concatenative synthesis is the use of diphones (Taylor, 2009). A diphone is a segment of two adjacent phones in which the beginning and ending of the diphone are the steady

state portions of each phone. Diphones are concatenated at these stable points where there is less spectral discontinuity between them (Taylor, 2009). This approach helps preserve the coarticulatory information and spectral transitions between phones — features that are difficult to model accurately in articulatory or formant-based synthesis methods. When a clearly articulated diphone inventory is used, the diphones can be concatenated with little or no smoothing at the transition, as the steady-state portions of the vowel will align (Taylor, 2009).

2.4.4. End-to-end Transformer Methods

Before deep learning approaches, speech synthesis relied on rule-based, parametric, and concatenative methods. These approaches divide the task into several stages which often demand language-specific, domain-specific, or otherwise customized modifications. Because these systems depend on rigid, sequential rules, they have fallen out of favor.

End-to-end deep learning approaches marked a transition in approach. Instead of splitting the process into independent, modular stages — each one a potential failure point — deep learning models are based on internal feature spaces within a unified network. They produce a speech waveform without discrete stages, which improves voice quality, speaker consistency, and adaptability (Shen et al., 2018). Such subtle features emerge from training rather than being explicitly modeled.

End-to-end transformer methods process input text through a sequence of neural layers that transform discrete symbolic inputs into acoustic output. This is performed by tokenizing inputs and mapping them to vector embeddings, with positional encodings applied to preserve word order information. These embeddings pass through multiple hidden layers that use multi-head self-attention to capture contextual relationships between input tokens, predicting phonetic and prosodic features such as stress, rhythm, and intonation. The attention mechanism aligns the symbolic text input with corresponding acoustic frames, producing a mel-spectrogram as the encoder-decoder output (Shen et al., 2018). This spectrogram is then passed to a neural vocoder — such as HiFi-GAN, WaveNet (van den Oord et al., 2016), or WaveGlow (Prenger et al., 2019) — which converts it into a final audio waveform. GAN-based vocoders such as HiFi-GAN are widely used for their efficiency and naturalness (Kong et al., 2020).

The biggest drawback of the deep learning approach is segmental and spectral limitations. Even subtle spectral features appear to be detectable across synthesis models, rather than being idiosyncrasies of particular models. Ustinov et al. (2025) demonstrated that a transformer-based detection model trained on one commercial synthesis model, ElevenLabs, can detect synthetic speech produced by other highly advanced models, including NotebookLM and Minimax. The ability to detect synthetic speech from models that the detection model was never trained on suggests that a common set of spectral anomalies distinguishes synthetic and human speech, regardless of the specific generative method employed. Ustinov et al. (2025) did not explore the specific common spectral artifacts that were common across synthesis models, leaving this for future research. However, this finding implies that these spectral divergences reflect more fundamental limitations in how current models mimic human speech without truly capturing the subtlety of articulatory and psycholinguistic demands.

These segmental and spectral limitations may also have bearing on the intelligibility of deep learning systems. This is demonstrated by Cohn and Zellou's (2020) comparative study in which they tested both neural and concatenative modes of Amazon's Polly (one of the models used in this thesis) under adverse listening conditions. Polly was prompted for speech in both modes, and the resultant speech was tested for word recognition by human subjects in a noisy environment. The results showed significantly degraded intelligibility for the more advanced neural mode relative to the concatenative mode, despite the neural mode receiving higher naturalness ratings. Cohn and Zellou (2020) attributed this to the neural mode's tendency toward more casual, connected-speech reduction patterns. Concatenative synthesis, on the other hand, benefited in intelligibility from its method of joining various segments of recorded clear speech, which served as an inadvertent proxy for the hyperarticulated end of the articulatory spectrum, and thus proved more robust in degraded listening environments. The implication is that neural speech synthesis models have learned to produce speech in a way that replicates human strategies of reducing articulatory effort and produce casual, natural-sounding speech under ideal conditions, but cannot adapt to more communicatively demanding situations. However, despite being less intelligible than concatenative models, under certain conditions, neural speech can still outperform human speech. Tånnander et al. (2023) demonstrated this using a multi-track method, in which multiple

simultaneous renditions of the same utterance were played to listeners. In this condition, neural speech proved more intelligible than human speech due to its lower variability across repetitions.

Deep learning models also lack the human ability to create spontaneous utterances. Ada et al. (2024) examined the degree to which current voice cloning tools could replicate paralinguistic phenomena, specifically stuttered speech. Their analysis revealed that when the speech synthesis model attempted to produce a disfluent repetition — for example, the word fragment in "consistent" with a stutter on the first phone — it failed to generate the reduced, centralized vowel that a human speaker would naturally produce during the disfluency ([k_]). Instead, the model substituted a full, unreduced vowel ([ko_]) drawn from a different lexical context, producing a sound more consistent with the initial syllable of an unrelated word such as *coordinate*. This finding suggests that the model lacks an internal representation of how vowel targets are modulated by the articulatory disruptions of natural disfluency. However, this type of issue is not unique to transformer approaches. Lorenzo-Trueba et al. (2019) trained a WaveRNN-based model on a dataset of 74 speakers across 17 languages and had human subjects judge the synthetic utterances. Though the model performed very well in clean conditions, its intelligibility dropped in noisy conditions, possibly indicating the same issue as was identified by Cohn and Zellou (2020). Most interestingly, Lorenzo-Trueba et al. (2019) noted specific shortcomings regarding non-speech vocalizations. While the model could handle sounds of disgust or pleasure, it struggled with high-energy bursts like grunts or shouts, often generating heavily distorted sounds. The resulting synthetic utterances exhibited a characteristic distortion and lacked the clean harmonic structure that the human larynx produces. The model's training data are primarily based on conversational speech, so generalizing beyond that to the full range of human vocal production remains a limitation of the neural approach.

The temporal organization of speech — the patterns of duration, rhythm, and prosodic prominence that structure utterances — arguably represents the domain in which synthetic speech diverges most conspicuously from human production. Human speakers modulate segmental durations, pause placement, and pitch contours in response to a vast array of factors, including syntactic structure, information status, emotional state, and the demands of the discourse context. Current speech synthesis models, while capable of producing acceptable prosodic contours in simple declarative sentences, struggle to replicate the fine-grained temporal control that characterizes natural speech in more complex communicative situations.

The difficulty of controlling temporal properties in synthetic speech extends to explicit attempts at manipulation through prompting. Ada et al. (2024) reported that instructing a voice cloning system to produce speech at a slower rate via a prompt such as "he said slowly" yielded inconsistent and often paradoxical results. In some instances, the system produced speech that was faster rather than slower, or it introduced unexpected pitch shifts, such as lowering fundamental frequency in conjunction with slowed tempo. In human speech, speaking rate and pitch are partially independent parameters that speakers control with considerable precision; the inability of current systems to decouple these dimensions reveals a lack of fine-grained control over the temporal and melodic structure of the output.

More severe anomalies can arise from failures in the attention mechanisms that underlie neural speech synthesis. Tånnander et al. (2023) documented some of these "attention errors," in which the alignment between the input text and the generated audio signal breaks down, producing speech that contains skipped, repeated, or missing speech segments. When generating large pools of renditions of the same sentence — in some cases up to 1,000 — these attention errors appeared among the outputs, apparently stemming from the inherent variability of neural speech synthesis (Tånnander et al., 2023). These are gross errors that no human speaker would produce, and they show the fragility of the attention mechanisms of transformers in synthesizing speech. A single textual input corresponds to a nearly infinite set of legitimate acoustic realizations, each differing in pitch, duration, emphasis, and emotional coloring. Human speakers typically navigate these possibilities with ease, selecting realizations that align with their communicative intent and are appropriate for the context. Speech synthesis models, by contrast, must bridge an enormous information gap between sparse textual input and richly specified acoustic output. When models default to averaged or prototypical realizations, the resulting speech lacks the variety and prosody of human speech (Raitio et al., 2020).

The evidence points to a consistent pattern: current neural speech synthesis models are most successful when producing speech that falls within a narrow band of articulatory effort and communicative demand — the casual, conversational register that dominates their training data. When the communicative situation calls for articulatory adjustments — e.g., greater clarity in noisy conditions or more precise segmental specification in ambiguous contexts — these systems reveal their limitations. Achieving truly human-like flexibility in articulatory effort modulation remains a fundamental and unresolved challenge for the field.

2.5. Speech Recognition

In its canonical form, speech recognition is the process of converting an acoustic speech signal into text (Ahlawat et al., 2025). Regardless of the specific system, the task involves transforming an input acoustic sequence $x = (x_1, x_2, \ldots, x_T)$, where x represents the input features and T is the length of the signal, into a corresponding output sequence $y = (y_1, y_2, \ldots, y_N)$ of textual tokens where N is the number of output tokens (Ahlawat et al., 2025).

The evolution of speech recognition methods closely parallels that of speech synthesis. As with speech synthesis, speech recognition has transitioned from modular, rule-based, and statistical approaches to end-to-end deep learning architectures. Before the emergence of end-to-end models, traditional speech recognition systems operated in several distinct stages. In the first stage — pre-processing — the acoustic signal is prepared through format conversion, filtering, and noise suppression to enhance speech.

In the second stage — feature extraction — coefficients are derived from the pre-processed signal using techniques such as Mel spectrograms and Mel-Frequency Cepstral Coefficients (Ahlawat et al., 2025). Mel spectrograms provide a time–frequency representation of the audio signal, and Mel-Frequency Cepstral Coefficients compress this representation by preserving only the frequency information most relevant to human perception.

The third stage involves classification, in which the extracted features are aligned with linguistic units (e.g., phonemes or characters) to generate text. In the final stage, a language model applies higher-level linguistic constraints, including grammar and semantic information, to refine and post-process the output text. One common post-processing algorithm is the n-gram model, which estimates the probability of a word based on the previous n words. For example, a bigram model ($n = 2$) predicts the next word pairwise based on the word before it. A language corpus is used to pre-compute the co-occurrence frequency of all bigrams, so that for any single word, the most likely bigram is selected based on the immediate context of n-1 (with the algorithm selecting *red wine* over *red whine*) (Jurafsky & Martin, 2023). Larger probability sets can be used where $n = 3, 4$, and so on. Such co-occurrence patterns represent the statistical regularities of a language that replaced the explicit rules of even earlier approaches (Jurafsky & Martin, 2023).

The n-gram model relies on the training data adequately covering real-world language use, as word sequences that are absent in the corpus receive zero probability regardless of their

plausibility (Brown et al., 1992). A complementary approach is the class-based model that looks at the part-of-speech or semantic grouping of a given word, and estimates the probability of the next word based on class. For example, the probability of a determiner (like *the*) being followed by another determiner is unlikely, whereas the probability of a determiner being followed by a noun is highly likely. By first classifying each word into a class and estimating the probability of class interaction, the model can generalize to unseen word combinations (Brown et al., 1992).

2.5.1. Hidden Markov Models

Due in part to their multidisciplinary appeal and effectiveness, HMMs were for a time the dominant speech recognition framework, and were widely used for several decades. The probabilistic state-transition nature of HMMs was well suited to recognizing acoustic speech signals (Rabiner, 1989). As described in 2.3, HMMs model language using a probabilistic sequence of hidden states and estimate the likelihood of observable acoustic features in the speech signal. Due to the continuous and inherently noisy nature of the speech signal, the mapping from the signal to linguistic units is probabilistic. Yet HMMs can still model underlying linguistic units through its sequence of hidden states with learned probability.

Since the Markov principle assumes that each state depends only on the immediately preceding state, this limited the model's ability to capture the longer-range phonological and syntactic dependencies that characterize natural speech (Gales & Young, 2008). Another major limitation of traditional HMMs in speech recognition is their inability to adequately capture the wide variability found in real-world speech. An HMM constrains how much a sound may vary across instances, so its acoustic modeling capacity is inherently limited. To address this problem, HMMs were soon combined with Gaussian Mixture Models (GMMs), which provide a more flexible estimate of the range of possible acoustic realizations under different speaking conditions by modeling each phoneme with multiple overlapping Gaussian distributions (Jurafsky & Martin, 2023). Although individual Gaussian functions had already been used with HMMs in some cases, GMMs generalize this approach by mixing several distributions to better represent variation in the data. The HMM–GMM hybrid architecture became the standard approach in speech recognition prior to the introduction of deep learning methods (Mehrish et al., 2023, Jurafsky & Martin, 2023).

However, HMMs still exhibited weaknesses in real-world speech conditions and degraded significantly in the presence of background noise or other conditions of high acoustic variability

(Jurafsky & Martin, 2023). These limitations established the conditions for the shift toward recurrent neural architectures, which offered a more flexible framework for modeling sequential dependencies in speech.

2.5.2. Recurrent Neural Networks

Since HMMs struggle to capture long-distance dependencies in speech, other modeling approaches emerged to overcome this limitation. RNNs were the next major step, shifting the field beyond purely statistical HMM-based methods into neural architectures. Unlike HMMs, RNNs preserve sequential context by feeding the hidden state from previous time steps back into the network. This recurrent mechanism enables the model to retain contextual information across longer spans of time, allowing it to capture temporal relationships more effectively than HMMs (Graves et al., 2013). A key development that made RNNs practical for speech recognition was the introduction of Connectionist Temporal Classification, a loss function that allowed RNNs to train directly on unsegmented acoustic sequences without requiring frame-level alignment between input and output, which had been a significant limitation of earlier HMM-based approaches (Graves et al., 2006). This enabled RNNs to learn the mapping from acoustic input to textual output in a more flexible and data-driven manner, improving speech recognition performance (Graves et al., 2013).

While RNNs do capture sequential contextual relationships better than HMMs, they do so in a limited fashion. Most significantly, the issue of vanishing and exploding gradients hampers their ability to retain memory over long distances (time steps), making the network less accurate on long sequences. This becomes apparent in speech recognition on longer utterances, and sometimes in conversational speech, where dependencies between distant phonetic and lexical elements are common (Jurafsky & Martin, 2023). Finally, due to the sequential nature of the processing, which includes recurrency, RNNs were inefficient on a large scale and specially made training computationally expensive, even with the constant increases in computing power. These limitations — gradient instability, poor long-range memory, and sequential processing constraints — led to the development of new types of RNNs, such as the LSTM, which introduced mechanisms specifically designed to address the shortcomings of standard RNNs.

2.5.3. Long Short-Term Memory Networks

As a more sophisticated version of an RNN, the LSTM was developed to address the shortcomings of RNNs (Hochreiter & Schmidhuber, 1997). The LSTM addressed the gradient problem by implementing several "gates" at each node in the network which regulate the flow of information through the cell and reduce gradient instability during training. The modification of the RNN neuron was achieved by adding specialized memory gates (forget, input, and output) to gate information and reduce vanishing and exploding gradients. This improved the ability of the network to accurately capture longer speech sequences, since the model was able to retain acoustic context across longer spans than standard RNNs (see Graves & Schmidhuber, 2005). However, despite these improvements, LSTMs still suffered from gradient issues, though less than standard RNNs. Graves et al. (2013) showed that deep bidirectional LSTMs trained with connectionist temporal classification achieved state-of-the-art phoneme recognition on the TIMIT benchmark, demonstrating the effectiveness of end-to-end recurrent neural architectures for speech recognition. As noted previously, training of RNNs and LSTMs is computationally expensive and slow since speech input is processed sequentially. The sequential nature of RNN processing created a computational bottleneck that limited the volume of training data that could be handled efficiently, constraining performance gains as datasets grew larger. LSTM usage thus remained constrained, and these limitations established the conditions for the shift toward transformer architectures, which eliminated recurrence in favor of parallelizable self-attention mechanisms.

2.5.4. End-to-end Transformer Methods

The multi-headed, attention-based transformer methods of Vaswani et al. (2017) eliminated the recurrence of RNNs and LSTMs and relied entirely on self-attention mechanisms. The use of massively parallel operations to capture long-distance (temporal) relationships also eliminated sequential processing, which significantly increased speed and throughput, allowing for vastly more training data to be used efficiently. Though Vaswani et al. (2017) implemented translation, this architecture was quickly adopted in speech models such as Dong et al. (2018). The transformer architecture was further augmented in Gulati et al. (2020) by adding convolutional layers, which captured both long-range dependencies and local acoustic features.

End-to-end, attention-based transformer models transform acoustic speech input directly to textual output without multiple intermediate stages, using the encoder and decoder mechanisms (Vaswani et al., 2017; Dong et al., 2018). The encoder converts the input sequence of spectral features into high-dimensional acoustic representations. The decoder then generates output text from the embeddings, predicting each token based on the current context and the previously generated tokens (Dong et al., 2018).

However, traditional attention mechanisms require access to the full input sequence, which limits their suitability for real-time use. While some approaches attempt to address latency issues, others focus on reducing the need for labeled data. One such approach is Meta's wav2vec 2.0, which Baevski et al. (2020) demonstrated can learn high-quality speech representations from vast amounts of unlabeled data during unsupervised training, making it more scalable in many applications. Rather than requiring transcribed speech for training, wav2vec 2.0 uses a contrastive learning objective to learn discrete speech units from raw waveforms, after which a relatively small amount of labeled data is sufficient to fine-tune the model for speech recognition. This approach substantially reduces the dependence on annotated corpora and has demonstrated strong performance across a range of languages and acoustic conditions, making it particularly attractive for low-resource settings (Baevski et al., 2020).

Radford et al. (2023) adopted a different approach with Whisper, training the model on approximately 680,000 hours of weakly supervised multilingual and multitask data collected from the web. Rather than relying on self-supervision, Whisper is trained as a sequence-to-sequence model that directly maps acoustic input to text across multiple tasks, including transcription, translation, and language identification. Its large training dataset enables better performance across domains, accents, noise conditions, and languages without task-specific fine-tuning (Radford et al., 2023).

While word error rate is traditionally used to measure speech recognition performance, recent studies have examined entropy and probability scores based on output-layer activations for additional purposes (Laptev & Ginsburg, 2022; Ravuri et al., 2024). While word error rate measures whether the model predicted the correct word, it does not indicate the network's confidence. Laptev & Ginsburg (2022) used entropy to calculate word-level confidence levels in correct and incorrect transcriptions in Conformer-CTC and Conformer-RNN-T models. Ravuri et al. (2024) showed that entropy measures from wav2vec and wav2vec 2.0 could be used to predict

human perceptual quality ratings (Mean Opinion Score) of synthesized speech. High entropy scores from the models correlated with low human mean opinion scores, allowing the models' scores to serve as proxies for human scores. This research illustrates the potential of using confidence and entropy to measure model confidence across varying speech input.

The use of confidence and entropy metrics to detect PND- and WF-related hyperarticulation and hypoarticulation in wav2vec 2.0, as in the present study, addresses an underexplored aspect of speech recognition research. Early evidence showed that hyperarticulated speech produced higher error rates in speech recognition systems (Butzberger et al., 1992). Since early speech recognition models were trained overwhelmingly on read speech, they encoded the statistical regularities of that speech style. When speakers hyperarticulated, the resulting acoustic signal deviated from the training data that the model expected. The hyperarticulation created spectral and temporal patterns that fell outside the model's training and expectations. This led to a counterproductive result — when a user encountered a misrecognition and responded by speaking more clearly, the hyperarticulation often worsened recognition further.

To mitigate the discrepancy between error rates in hyper-articulated and neutral, or hypoarticulated, speech, data augmentation techniques have been employed to target the acoustic-phonetic variability inherent in various speech patterns. Lee et al. (2018) synthetically generated both hyperarticulated and hypoarticulated speech from the neutral training corpora, in lieu of naturally occurring hyperarticulated and hypoarticulated speech. To augment the training data, the authors employed time-scale modification (TSM) to programmatically alter speaking rates — accelerating and decelerating the audio to simulate the hypoarticulation and hyperarticulation, respectively. Their techniques also included manipulating included manipulating formant bandwidths and creating spectral flattening, features associated with hyperarticulation. By expanding the acoustic model's exposure to these synthetically generated styles — which mimic hyper- and hypoarticulated speech — Lee et al. achieved performance gains in speech recognition of both hypoarticulated and hyperarticulated utterances.

Transformer-based speech recognition models represent a substantial advance over their RNN and HMM predecessors, offering greater scalability and flexibility across diverse acoustic conditions and languages. Nevertheless, challenges remain in handling the full range of natural speech variability, including dialectal variation, spontaneous speech, and articulatory extremes, which continue to drive active research in the field.

Chapter 3: Methodology

The objective of this study is to examine how phonological neighborhood density (PND) and word frequency (WF) influence the acoustic properties of Polish vowels, and if comparable effects exist in AI-based speech synthesis and recognition systems. This study extends prior work on PND and WF effects — predominantly conducted on English — to Polish, an underexplored language in both psycholinguistic and natural language processing contexts. This study is the first systematic comparison of these effects across human and AI speech modalities in Polish.

Throughout this study, non-random variations in speech production — such as the effects of PND and WF — are referred to collectively as systematic articulatory variability (SAV). SAV denotes the statistically emergent adjustments of articulatory targets in response to phonological, lexical, and contextual properties of words. This term is used to distinguish these effects from random phonetic variation, and to provide a unified approach for comparing human and artificial speech behavior.

This study addresses three key research questions:

1. **Human speech data:** How do PND and WF influence the acoustic realization of vowels in Polish? Specifically, are words in dense phonological neighborhoods with low lexical frequency (‘hard’ words) characterized by hyperarticulation, while words in sparse neighborhoods with high lexical frequency (‘easy’ words) show hypoarticulation?
2. **AI speech data:** In Polish, to what extent do transformer-based speech synthesis models reproduce the PND- and WF-mediated effects found in the human vowel space?
3. **AI recognition data:** In Polish, do the internal representations of words within a transformer-based speech recognition model reflect a structure analogous to human vowel space? In particular, do ‘hard’ and ‘easy’ words occupy distinct regions within the model’s representational space, corresponding with their acoustic differentiation in human speech, and do confidence metrics vary by difficulty condition like human perception?

To address these research questions, a controlled set of monosyllabic Polish words along a continuum from hard to easy was used as stimuli to solicit three types of data: (a) vowel production by native Polish speakers, (b) vowel production by speech synthesis systems, and (c) internal activation within a speech recognition model. This multi-modal approach enabled a comparison between native-speaker production, synthesized speech, and artificial perception within a single study.

3.1. Data Stimuli

The stimulus set consisted of 104 monosyllabic Polish words. All items were consonant-vowel-consonant (C^+VC^+) structures to control for syllable complexity. Each word contained one of the eight Polish vowels (/a, ɛ, i, ɨ, o, u, ɔ̃, ɛ̃/).

Out of the 104 tokens, 30 were classified as hard, 23 were classified as easy, and 51 were in the control group. Only words containing one of the six oral vowels were examined, while those with nasal vowels were included as distractors and excluded from the data analysis. Levenshtein neighborhood density and frequency-per-million values were obtained from the Jiwar database (Alzahrani, 2025) to construct the dataset. Stimuli were selected from all six vowel categories to cover the four combinations of PND and WF values. Words were ranked by their combined PND and WF properties. Words with the highest PND and lowest WF were classified as hard (n = 30), and words with the lowest PND and highest WF were classified as easy (n = 23). This corresponds to 51% of the data set, of which 29% (n = 30) were hard words, and 22% (n = 23) were easy words. The phonotactics of Polish constrained the selection of the data set in obtaining an even split between categories, given the number of monosyllabic C^+VC^+ words in the lexicon unequally distributed across the six oral vowels and difficulty categories. The remaining 49% of the dataset (51 words) in the PND and WF distributions, which included nasal vowels, served as distractors. The distributions of PND and WF values across the stimulus set are listed in the appendix.

Words were selected using a two-factor design representing high-versus-low PND and high-versus-low WF, producing four lexical categories (summarized in Figure 3.1):

1. Dense neighborhood + low-frequency = ‘hard’ words

2. Sparse neighborhood + high-frequency = ‘easy’ words

3. Dense neighborhood + high-frequency = mixed effects

4. Sparse neighborhood + low-frequency = mixed effects

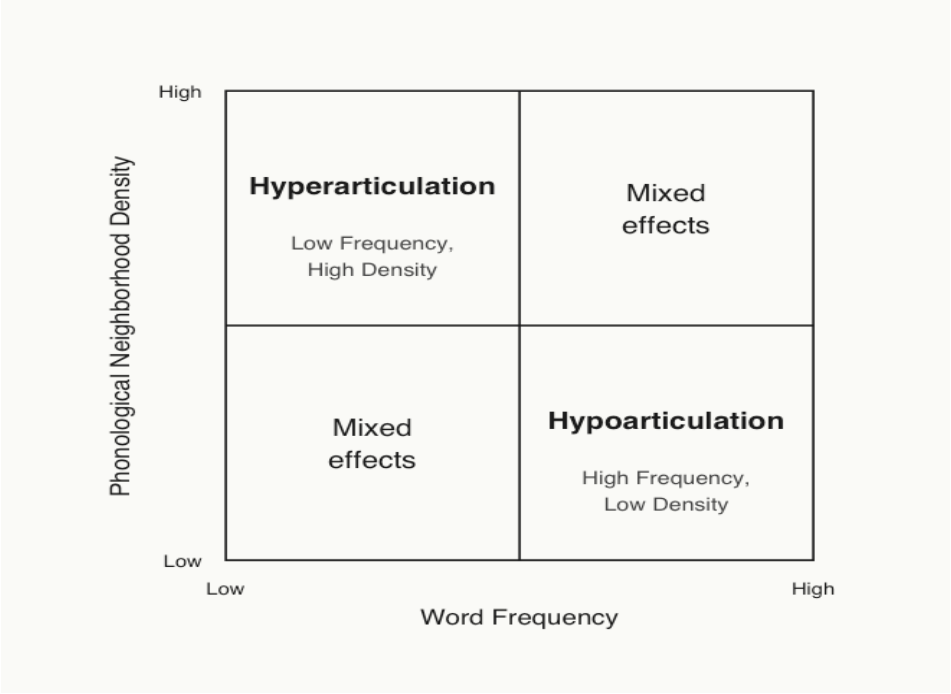

Figure 3.1. Combined acoustic effects of PND and WF (author-created).

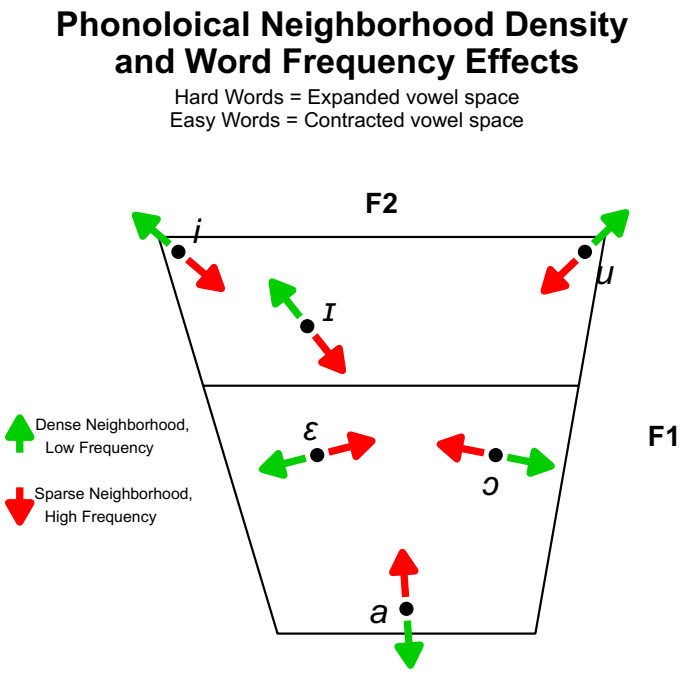

Figure 3.2. Vowel space effects predicted by PND and WF studies for Polish (author-created).

While the vast majority of the stimuli were monophthongs, four tokens contained glides: two diphthongs (*maj*, *spójrz*) and two with onset-initial glides (*głos*, *jeść*). The decision to retain these tokens warrants discussion, as previous studies have adopted different approaches. Unlike monophthongs, which exhibit relatively steady-state formant values amenable to measurement at a single temporal point, diphthongs and glides involve continuous formant movement throughout the vowel, impeding the identification of a stable acoustic target. This dynamic quality introduces decisions about where to sample — onset, midpoint, or offset — and adds a degree of complexity that could obscure the effect under investigation. Munson and Solomon (2004), Stephenson (2004), and Gahl et al. (2012) excluded diphthongs, whereas Wright (2004) retained them.

In this study, diphthongs were retained for two reasons. First, the small number of monosyllabic Polish $C^{+}VC^{+}$ words with suitable PND and WF distributions limited the available stimulus pool. Therefore, retaining stimuli with diphthongs allowed for a larger and more balanced

set. Second, their retention tests whether PND and WF effects extend to vowels with dynamic formant trajectories rather than applying only to steady-state monophthongs. Midpoint formant extraction provides a consistent measurement point that still captures the vowel's target, despite the influence of the formant movement of diphthongs. This approach allowed the diphthongs to be treated the same as monophthongs during the Euclidean distance calculations. Displacement of the midpoint measurements for the four tokens with glides is reported in Chapters 4 and 5.

3.2. Experimental Design

The experiment required three data types: human speech data, synthesized speech data generated by two text-to-speech systems, and measurements of activations of a speech recognition model. All components relied on the same 104-word stimulus list for consistency across production, synthesis, and recognition tasks. In the human production and speech synthesis data, each vowel segment was isolated, annotated, and subjected to acoustic analysis. The speech recognition data was obtained via internal network activations in response to the human speech recordings. These activations were then analyzed as a function of lexical characteristics.

The design followed methodological standards established in related research, including the easy/hard stimulus design (Bradlow & Pisoni, 1999), inter- and intra-vowel distance methodology (Stephenson, 2004), AI-synthesis comparison protocols (Song et al., 2025), and wav2vec 2.0 internal representation analysis (tom Dieck et al., 2022).

3.2.1. Human Participants

A total of 26 native speakers of Polish participated in the experiment. All speakers were adults aged 19-22 (M = 19.77), with no reported history of speech, language, or hearing disorders. Participation was voluntary, and speakers were recruited through the university. Of the 26 participants, 20 were female, and 6 were male. This gender imbalance reflects the demographic makeup of the pedagogical program at the university from which participants were recruited, in which female students make up the majority. Lobanov normalization was applied to formant values to control for speaker-level variation. All speakers self-reported as native speakers of standard Polish with no regional dialects. All participants provided informed consent, and data collection procedures adhered to the university's ethical guidelines for human participants research in linguistics.

3.2.2. Speech Synthesis Model 1 (Amazon Polly)

Amazon's neural speech synthesis model, Polly, which employs sequence-to-sequence modeling with attention (Amazon Web Services, 2024), was selected in part due to its role in prior research on lexical frequency effects on speech synthesis in English (Song et al., 2025), allowing for direct cross-linguistic comparison of results. At the time of data collection, only two Polish voices (Ewa and Ola, both female) were available in Amazon Polly's neural engine. Although lacking in gender diversity, these two voices are representative of Amazon's Polish neural speech.

3.2.3. Speech Synthesis Model 2 (ElevenLabs)

Speech was also elicited from two synthetic Polish voices from ElevenLabs. The ElevenLabs model was selected due to its popularity and industry-leading role in artificial speech synthesis. Although the details of its architecture are also proprietary, ElevenLabs company documentation states that its model uses neural acoustic modeling. The process for accessing speech generated through ElevenLabs differs from Amazon's approach, in that Amazon offers only a set number of 'voices,' whereas ElevenLabs allows users to create their own custom voices. This creates a complication for research design, as it is not possible to obtain Polish synthetic voices directly comparable to Polly's Polish-trained voices. To address this and to introduce gender balance — which was not possible with Amazon's voices, since both Polish neural voices are female — both male and female voices were generated. This research design allowed for a preliminary empirical test of whether such speech synthesis systems could approximate the human vowel-space dynamics of PND and WF.

3.2.4. Artificial Speech Recognition Model (wav2vec 2.0)

To evaluate possible perceptual effects of lexical competition in an artificial recognition system, human and synthetic tokens were processed through wav2vec 2.0 (Baevski et al., 2020). Activations were extracted from the wav2vec2-large-xlsr-53-polish model, which is a variant of wav2vec2-large consisting of 24 transformer layers, pretrained on 53 languages (Conneau et al.,

2021) and subsequently fine-tuned on Polish using Mozilla Common Voice 6.1[1]. All .wav files were provided to the model one-at-a-time for sequential processing. The speech recognition analysis did not require Praat textgrids, as vowel identification and alignment were performed by wav2vec 2.0. Data processing and extraction from wav2vec 2.0 were implemented in Python using the `torch` and `transformers` libraries.

To accomplish the PCA step, activations were extracted from Layer 24 of the transformer, which prior work has shown contains phonetically interpretable representations suitable for PCA projection in F1/F2 space (tom Dieck et al., 2022). To assess the model's confidence, softmax probabilities were computed (using the `PyTorch` library) from the output logits of the classification head applied to the final transformer layer's activations. Confidence values were measured from the softmax probability for each predicted phoneme from the entire phoneme vocabulary of wav2vec 2.0. Shannon entropy (using the SciPy library) was also computed over the full probability distribution to indicate uncertainty. Higher maximum probability was interpreted as greater confidence of the vowel selected by the network, and higher entropy indicated a wider distribution across competing vowels (and other phonemes) and therefore lower confidence. Both measures were computed for each target vowel for subsequent analysis.

3.3. Data Collection

Recordings of native speakers were obtained in the Speech Processing Laboratory at the University of Silesia in Katowice. Stimuli were presented in isolation to avoid prosodic or semantic contextual effects, adapted from established methodologies (see Munson & Solomon, 2004; Scarborough & Zellou, 2013). Stimuli were presented visually on a computer screen in three randomized sets using a Microsoft PowerPoint slide deck. The stimuli for each of the three sets were independently randomized to mitigate order effects. Participants were instructed to read each word naturally, as they would in everyday speech. Speech tokens exhibiting disfluency, such as stuttering or misread words, were excluded. During acoustic analysis, tokens with audible microphone artifacts (signal clipping or off-axis recording) were also excluded. Signals were

[1] The Polish fine-tuned model was released by Grosman (2021) and is publicly available at https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-polish. The model was fine-tuned on Common Voice 6.1 (see Ardila et al., 2020, for a description of the Common Voice dataset).

recorded directly into a Sennheiser HMD 26 dynamic microphone headset, preamplified and digitized via a Sound Devices USBPre2 audio interface, and saved in uncompressed .wav format at 44.1 kHz, 16-bit, mono. These specifications provided sufficient temporal and spectral resolution for accurate measurements (e.g., formant trajectories, amplitude, duration).

The native-speaker data contained 8,112 tokens (26 speakers × 104 stimuli × 3 repetitions), and the speech synthesis data contained 1,248 tokens (4 synthetic voices × 104 stimuli × 3 repetitions), for a total of 9,360 tokens. Tokens were stored in a structured manner with metadata, including speaker ID, vowel category, PND values, WF values, and repetition number. These measurements served as the basis for all statistical comparisons between lexical conditions.

For the AI-generated data, productions were elicited and downloaded in .wav format through the Amazon and ElevenLabs web interfaces to maintain consistency with the human data. While every effort was made to replicate the human experiment using AI speech synthesis, a different prompting method was required, since it is not possible to present the AI model with data stimuli in the form of PowerPoint presentations. Initial efforts to elicit utterances by presenting a list to the AI models proved ineffective due to irregularities in responses, including mispronunciations, skipped words, and other inconsistencies. This is similar to methodological issues encountered by Tånnander et al. (2023). In response, the methodology was updated to use a carrier sentence in the form of “Mówię słowo [target word] po polsku” (“I say the word [target word] in Polish”).

For the speech recognition activation data, each .wav file was programmatically fed into a bespoke Python program, which was processed by the local copy of wav2vec 2.0, version wav2vec2-large-xlsr-53-polish. Only human speech was used to test wav2vec 2.0 since synthetic speech does not consistently demonstrate PND and WF effects. The data were collected via a programmatic inspection of the internal activations using the same Python script. This yielded two sets of speech recognition data. The first set included the network’s internal PCA embeddings as the model’s representation of a vowel chart (tom Dieck et al., 2022). The second set included two activation values — maximum probability and entropy — which, when combined, indicate the model’s confidence in the vowel classification (Ravuri et al., 2024).

3.4. Data Processing

All speech tokens were segmented and annotated using Praat version 6.3.09 (Boersma & Weenink, 2023). Vowel onset and offset were marked using both waveform and spectrographic cues, with manual correction where necessary to ensure labeling accuracy. Formant values at the midpoint of the vowel (see Figure 3.3) were extracted using the Parselmouth library (Jadoul et al., 2018) interface to Praat in Python. Unusual or outlier formant values, particularly spurious F1/F2 tracking errors, were identified through visual inspection of formants and excluded from the data. Tokens failing quality thresholds were excluded from the study, consistent with procedures used in comparable vowel-space studies.

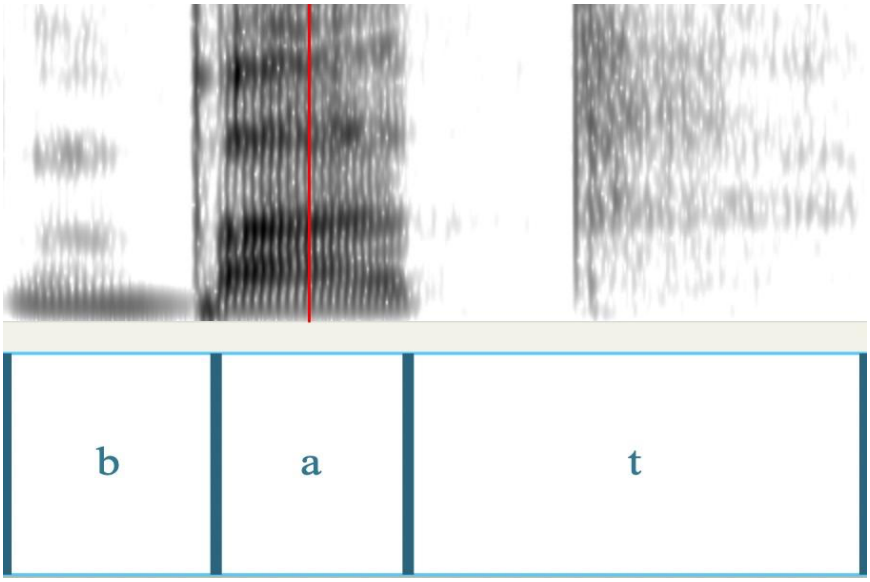

Figure 3.3. Spectrogram of the Polish word *bat* with the midpoint between vowel onset and offset shown in red.

To control for differences between speakers, formant values were normalized using the Lobanov method, then converted from Hz to the Bark scale. This conversion was applied to ensure that distance measurements reflect perceptually meaningful differences in vowel quality rather than raw acoustic frequency differences, which are not linearly related to perceived distance in human perception. All centroid-based, inter-vowel, and intra-vowel Euclidean distance measures reported in Chapter 4 were computed in Bark-scaled F1–F2 space.

3.4.1. Centroid-based Analysis

This analysis defined the degree of expansion/contraction from the centroid by measuring Euclidean distances from the vowel midpoint of each token — using the F1 and F2 values in two-

dimensional space on the *y* and *x* axes, respectively — to the vowel space center for speakers. While this distance is the primary measurement of vowel hyper- and hypoarticulation in hard and easy words, two other supplementary measures were used.

3.4.2. Inter-vowel Analysis

Inter-vowel distance was calculated following the methodology of Stephenson (2004) as a complementary measurement of vowel-space dispersion. For each token within a given difficulty condition (hard or easy), the mean Euclidean distance to every other token outside its vowel category is calculated, as shown in Figure 3.4. These distances are then averaged by difficulty condition to quantify the effects of PND and WF. Unlike the centroid-based measure, which quantifies expansion from a global vowel-space center, inter-vowel distance is more sensitive to system-wide effects (Stephenson, 2004).

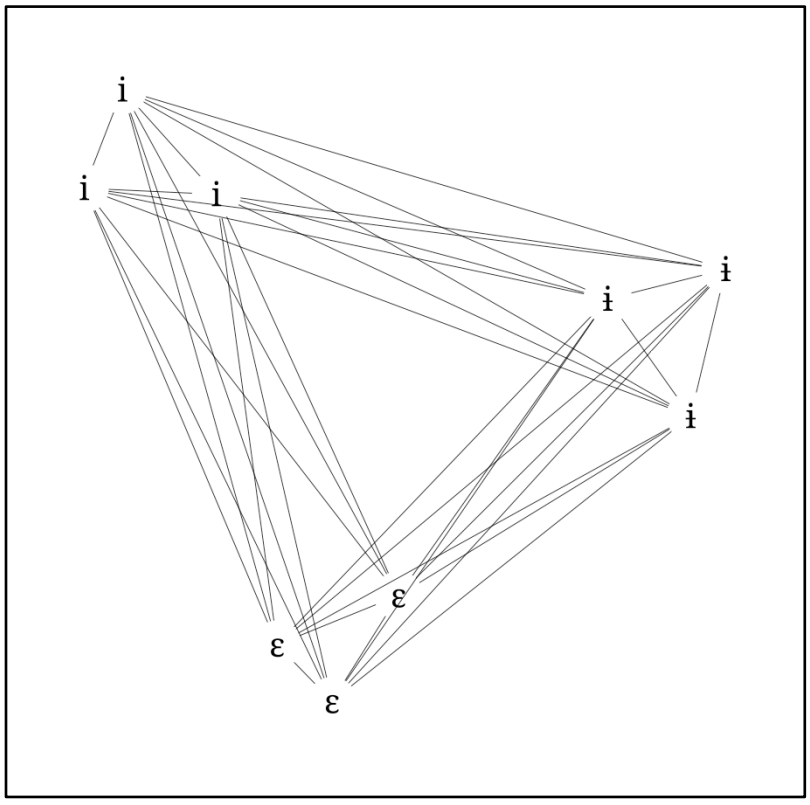

Figure 3.4. Inter-vowel distance (adapted from Stephenson (2004))

3.4.3. Intra-vowel Analysis

Intra-vowel distance was also calculated following the methodology of Stephenson (2004) as a measurement of vowel token clustering. For each vowel, the mean Euclidean distance between all tokens within that vowel category was computed and averaged by difficulty condition. Unlike inter-vowel distance, which measures separation between categories, intra-vowel distance reflects how tightly tokens are grouped within a given vowel and difficulty condition. This measurement indicates articulatory precision of the vowels within their category. Larger intra-vowel distances indicate greater dispersion among tokens, suggesting lower articulatory precision. Conversely, smaller distances reflect tighter clustering of tokens in the acoustic space, indicating more stable and precise vowel production.

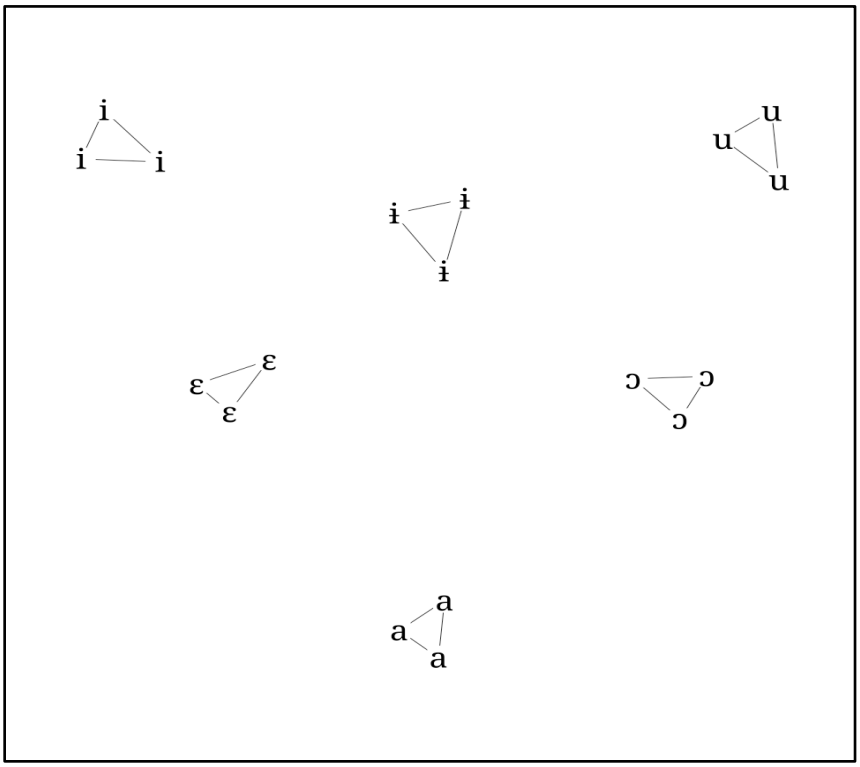

Figure 3.5. Intra-vowel distance (adapted from Stephenson (2004)).

3.4.4. Angle of Deviation

Finally, the comparative nature of this study, which contrasted human and AI speech within a single framework, permitted an exploratory measure that was not anticipated in the original research design. The angle subtended at the centroid between the averaged easy and hard plots for each vowel — formed by their respective radial vectors — was calculated and compared between

human and synthetic speech. Because prior studies on PND and WF have examined human speech almost exclusively, angular deviation between hard and easy vowel variants has not previously been explored; Song et al. (2025), who studied PND and WF effects in speech synthesis and recognition, did not examine this measure either. In human data, the angle of deviation, based the current study, is minimal, so there was likely no impetus to investigate the issue further in previous studies, which may have prevented studying this aspect in AI-generated speech. This measure is therefore treated as a finding unique to the human-AI comparison and is reported in this study as an exploratory metric.

3.5. Artificial Speech Recognition Data Processing

Data were extracted from two locations in the wav2vec 2.0 network: the final transformer layer, to assess whether the network's encoding of the hard/easy contrast most resembles human acoustic representation, and the output layer, to examine whether the network's confidence differs between hard and easy words.

3.5.1. Vowel Chart Extraction at the Transformer Level

The hidden layer extraction methodology was adapted from that of tom Dieck et al. (2022). Internal representations of wav2vec 2.0 encode phonetic information in a structure analogous to a formant-based vowel space. While tom Dieck et al. fine-tuned wav2vec 2.0 base, the present study used a publicly available variant, wav2vec2-large-xlsr-53-polish (Grosman, 2021), which was already fine-tuned on Polish using Mozilla Common Voice 6.1 (Ardila et al., 2020) on top of an XLSR-53 multilingual base (Conneau et al., 2021). Therefore, no further fine-tuning was required.

To visualize the structure of the model's internal representations, following tom Dieck et al. (2022), Principal Component Analysis (PCA) was applied to the 768-dimensional hidden-layer activations extracted from the final transformer layer. PCA reduced the high-dimensional activation vectors to a two-dimensional space, allowing direct comparison with the traditional F1–F2 vowel chart. This dimensionality reduction was used to assess two variables: whether the model's learned representations resembled the human vowel chart, and whether tokens from each difficulty condition occupied distinct areas in the compressed representational space. The resulting two-dimensional PCA projections were examined for correspondence with human and AI vowel

space. This comparison was performed via the relative positions of vowels in the vowel space of the other two data types.

3.5.2. Confidence Activation Values at the Output Level

To assess confidence, the methodology of Ravuri et al. (2024) was adapted to use the softmax probability of the predicted phoneme to examine whether the model's confidence level at the phoneme classification stage correlates with psycholinguistic difficulty, as reflected by PND and WF characteristics. The maximum probability and entropy values were extracted from the output layer of wav2vec 2.0 and measured against the token difficulty condition. This methodology was chosen over word error rate (WER) and other traditional methods of evaluating speech recognition models to obtain a more nuanced measurement of the impact of difficulty conditions on the model's output. Entropy and maximum probability were treated as complementary measures of recognition certainty. High maximum probability combined with low entropy is interpreted as high model confidence. Conversely, low maximum probability combined with high entropy (i.e., a flatter distribution of the probabilities) reflects uncertainty.

3.6. Statistical Significance

LME models tested the overall main effect of lexical difficulty across all vowels. Per-vowel pairwise comparisons used Welch two-sample t-tests. The pairwise comparisons were conducted across six vowel categories (/a, ɛ, i, ɨ, ɔ, u/) for three types of measurements: centroid-based, inter-vowel, and intra-vowel. The AI speech recognition data also underwent calculation of speech recognition confidence by difficulty condition across the same six vowel categories (/a, ɛ, i, ɨ, ɔ, u/), plus Welch t-tests.

This study was conducted in R using RStudio (version 2023.09.0+463). Two complementary analyses were performed for each of the three acoustic measurements (centroid-based distance, inter-vowel distance, and intra-vowel distance). Linear mixed-effects (LME) models tested the overall main effect of lexical difficulty across all vowels, with random intercepts for speaker and for vowel nested within speaker. Because formant values were Lobanov-normalized prior to analysis, between-speaker variance in the LME was minimal. This analysis was applied to both human production data and the AI synthesis data. Welch t-tests were performed

using the R `t.test()` function, assuming unequal variances (`var.equal = FALSE`). The analysis used the `lme()` function from the `nlme` package.

The different measures required different levels of data aggregation before analysis. Centroid-based Euclidean distance was analyzed using individual vowel target tokens (3,173 total for the human dataset). Because inter-vowel and intra-vowel distances are computed by comparing groups of tokens against one another, these analyses used averaged values rather than individual tokens. Each speaker's vowel production was averaged into one value per vowel, per difficulty condition, producing a smaller analytical dataset. The speech-recognition confidence and entropy measures from wav2vec 2.0 were analyzed at the token level. The degrees of freedom reported for each test in Chapter 4 reflect these differences in analytical units.

Per-vowel pairwise comparisons of production data used Welch t-tests across the six oral vowel categories (/a, ɛ, i, ɨ, o, u/). Mean differences were computed as hard − easy, such that positive values indicate greater values in the hard condition. A Bonferroni correction for six vowel comparisons yielded an adjusted significance threshold of $\alpha = .0083$ (.05 / 6), which was applied across all three distance measurements and the two AI recognition confidence measures.

For the AI speech-recognition data, the effect of lexical difficulty was evaluated at the output level using two complementary confidence measures: maximum softmax probability and Shannon entropy. These were analyzed using the same approach (LME for overall main effect, Welch t-tests with Bonferroni correction for per-vowel comparisons), with separate LME models for each measure.

3.7. Conclusion

This chapter has outlined a three-part methodological framework for the comparison of PND and WF effects in AI and human speech, and AI speech recognition, in Polish. Human production, AI production, and AI recognition data were all obtained from the same stimulus set of 104 monosyllabic C^+VC^+ words across various (hard, easy, and two mixed conditions).

The human production data were obtained from 26 native Polish speakers at the University of Silesia in Katowice via randomized repetitions of the 104-word stimulus list, yielding 8,112 tokens. The AI speech synthesis component of the study used the same stimuli and acoustic analysis procedures as the human group. Output was generated by two neural text-to-speech

systems: Amazon's Polly and ElevenLabs. Two Polish voices from each system (two female in Amazon's model, and one female and one male in ElevenLabs) produced a total of 1,248 synthetic tokens across three repetitions of the 104-word stimulus list. The speech samples were subjected to the same Python segmentation using Praat libraries, formant extraction, normalization, and statistical procedures. This design allowed direct acoustic comparison between human and synthetic vowel-space behavior, and enabled evaluation of whether current neural-based speech synthesis models reproduce the lexically conditioned articulatory variations that are characteristic of human SAV.

F1 and F2 values were extracted at the midpoint of each vowel. All samples were normalized using the Lobanov method to account for speaker differences and converted to the Bark scale. The centroid-based Euclidean distance and inter-vowel distance measured vowel hyper- and hypoarticulation, while intra-vowel distance and angular deviation measured directional consistency in vowel space. Statistical analysis employed Welch t-test and linear mixed-effects models with Bonferroni correction across six pairwise vowel comparisons.

The AI speech recognition component of the study used Meta's Polish wav2vec 2.0 variant to assess whether PND and WF-related acoustic properties are detectable in the model's internal representations and output layer. First, PCA was applied to the 768-dimensional hidden-layer activations from the final transformer layer to assess the model's compressed representational space as an artificial corollary of vowel space, and to examine whether hard and easy word tokens occupy distinct regions within it. Second, maximum probability and entropy values were extracted from the model's output layer and treated as complementary indices of confidence.

These components form a consistent methodology in which a unified stimulus set, acoustic procedures, and statistical approach are applied across human production, synthetic production, and synthetic recognition to permit direct comparison of the appearance of SAV across all three modalities, and provide the basis for the quantitative results reported in Chapter 4.

Chapter 4: Analysis and Results

Since the effects of phonological neighborhood density (PND) and word frequency (WF) in Polish remain an open question, these results first explore the degree to which the expected phenomena are present in native Polish speakers. Given that vowel-space expansion and contraction based on PND and WF are explained by theories of human neural competition, articulatory effort, and strategies to maximize listener comprehension, the other primary focus of this inquiry is the degree to which these phenomena are present in synthetic speech. If the speech of AI models truly mimics human speech, then it should follow similar patterns in vowel space expansion and contraction, which could indicate that it has adapted to subtle human speech patterns and formed an internal representation of lexical effects. To date, these phenomena are severely understudied in AI-generated speech; this research therefore seeks to address this research gap in both Polish and AI-generated speech.

4.1. Measurements

For each speech production dataset (native Polish speakers and AI-generated speech), the distribution of vowels was calculated by extracting the F1 and F2 formants at the midpoint to arrive at the classic representation of the vowel in vowel space. These values were then averaged for each instance of a given vowel in both the hard-word and easy-word categories to show the net discrepancy between the two. These pairs of averaged vowel targets were plotted to visualize differences in vowel pronunciation and the directionality of those differences between difficulty conditions. To estimate the Euclidean distance of each vowel from the center, the average of all tokens' (including distractors) midpoints was used to compute a global centroid of the vowel space. Lobanov normalization was used to account for differences among individual speakers.

Several tokens with glides — *głos*, *maj*, *jeść*, and *spójrz* — were analyzed with extra scrutiny due to their coarticulatory influence on vowel midpoint measurements, which affected their positions in vowel space. In *głos*, the labio-velar glide /w/ produces a continuous formant transition into the vowel without the clean boundary of a consonant, shifting the measured midpoint toward the /w/ closure in the direction of /u/ in the high back region of the vowel space.

In *maj*, *jeść*, and *spójrz*, the palatal glide /j/ creates a fronting and raising effect on the adjacent vowel target, shifting the midpoint of the vowel measurement toward the high front region of the vowel space. These tokens were flagged prior to analysis. The inclusion of these tokens and their effect on the significance of the results are discussed in Chapter 5.

4.2. Human Speech Results

Research Question 1 (Human Speech Production): How do PND and WF influence the acoustic realization of vowels in Polish? Specifically, are words in dense phonological neighborhoods with low lexical frequency ('hard' words) characterized by hyperarticulation, while words in sparse neighborhoods with high lexical frequency ('easy' words) show hypoarticulation?

The results from the human data revealed vowel expansion in hard words (dense PND/low WF) relative to easy words (sparse PND/high WF), meaning that vowels in hard words were hyperarticulated compared to vowels in easy words. In the data, vowels of hard words, on average, consistently appear further from the centroid than those of easy words, as shown in Figure 4.1. According to these results, Polish exhibits hyper- and hypoarticulatory patterns similar to those of previously studied languages (Wright, 2004; Scarborough & Zellou, 2013; Scarborough et al., 2018), which supports the theory that PND and WF belong to the category of broad, cross-linguistic drivers of SAV.

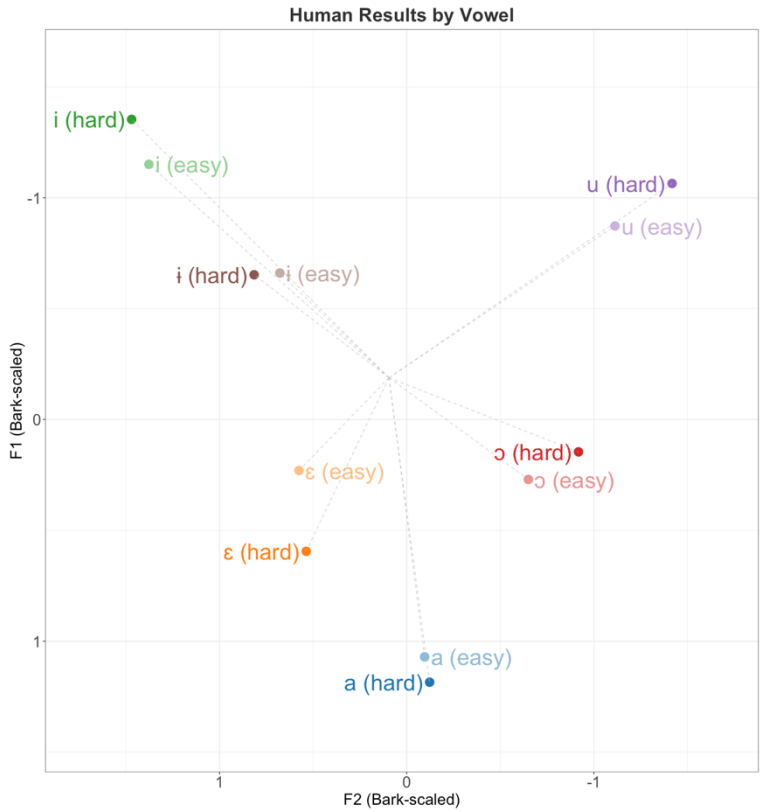

Figure 4.1. Human results in vowel space.

4.3. Statistical Significance of Euclidean Distance from the Centroid

The LME model revealed a significant main effect of lexical difficulty on Euclidean distance (β = .164, 95% CI [.139, .189], $t(3016) = 13.02$, $p < .001$), confirming that the expected effect is present in Polish human speech. The results and the associated statistical analysis (see Table 4.1) show that for all six vowels, hard words were produced with greater Euclidean vowel distance than easy words. The strongest effect was found for /i/, with a difference in Euclidean distance of .208, followed by /ɛ/ (*MD* = .204) and /u/ (*MD* = .168). The vowel /ɨ/ (*MD* = .085) also showed a smaller but still statistically significant expansion effect. Overall, these results demonstrate that lexical difficulty is meaningfully associated with vowel hyperarticulation.

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/i/	.208	5.495	209.58	< .05*
/a/	.155	6.775	545.06	< .01**
/ɛ/	.204	7.967	557.13	< .01**
/ɔ/	.146	4.728	325.58	< .01**
/u/	.168	3.894	553.26	< .01**
/ɨ/	.085	3.249	259.25	< .05*

Table 4.1. Human results: per-vowel Welch t-tests comparing hard and easy words in Euclidean distance from the vowel-space centroid, expressed in Bark. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

The data also revealed a significant effect of lexical difficulty with regard to inter-vowel distance (β = .155, 95% CI [.132, .178], $t(148) = 13.42$, $p < .001$), with hard words producing greater systemwide dispersion in vowel space than easy words, consistent with studies in English (Stephenson, 2004). Pairwise comparisons confirmed significant hard-easy differences for five of the six Polish vowels in human subjects, with /ɔ/ being the sole exception, as shown in Table 4.2.

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	.251	11.102	46.29	< .05*
/ɛ/	.084	3.696	40.43	< .01**
/i/	.27	5.174	46.24	< .01**
/ɔ/	.064	1.855	44.34	.07
/u/	.145	4.328	46.98	< .01**
/ɨ/	.104	3.324	47.89	< .05*

Table 4.2. Human results: per-vowel Welch t-tests comparing hard and easy words in inter-vowel distance, expressed in Bark. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

Stephenson (2004) reported no pattern of intra-vowel dispersion across vowel categories, and the Polish data also mirror this outcome. The LME model showed no significant overall effect of lexical difficulty on intra-vowel distance (β = −.041, 95% CI [−.101, .019], t(148) = −1.356, $p = .177$), with only /u/ (Table 4.3) reaching significance in the reversed direction (β = −.275). These

results suggest that, in Polish as in English, intra-vowel dispersion is not affected by lexical difficulty.

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	.147	1.679	46.78	.100
/ɛ/	-.036	-.463	27.93	.647
/i/	-.206	-1.427	47.79	.160
/ɔ/	.099	.794	41.22	.432
/u/	-.275	-2.849	45.25	< .05*
/ɨ/	.005	.108	45.09	.914

Table 4.3. Human results: per-vowel Welch t-tests comparing hard and easy words in intra-vowel distance, expressed in Bark. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

4.4. AI Speech Synthesis Results

Research Question 2 (AI Speech Synthesis): In Polish, to what extent do transformer-based speech synthesis models reproduce the PND- and WF-mediated effects found in the human vowel space?

In contrast to the results of native Polish speakers, the AI-generated speech revealed modest vowel-space expansion or contraction, and one vowel even showed the opposite effect to that observed in the human results. After the Bonferroni correction was applied, the aggregated AI data showed significant hard-easy differences only for /a/ and /i/, with the remaining four vowels failing to do so in pairwise comparisons (see Table 4.4). The average vowel plots for each vowel by difficulty category are illustrated in Figure 4.2.

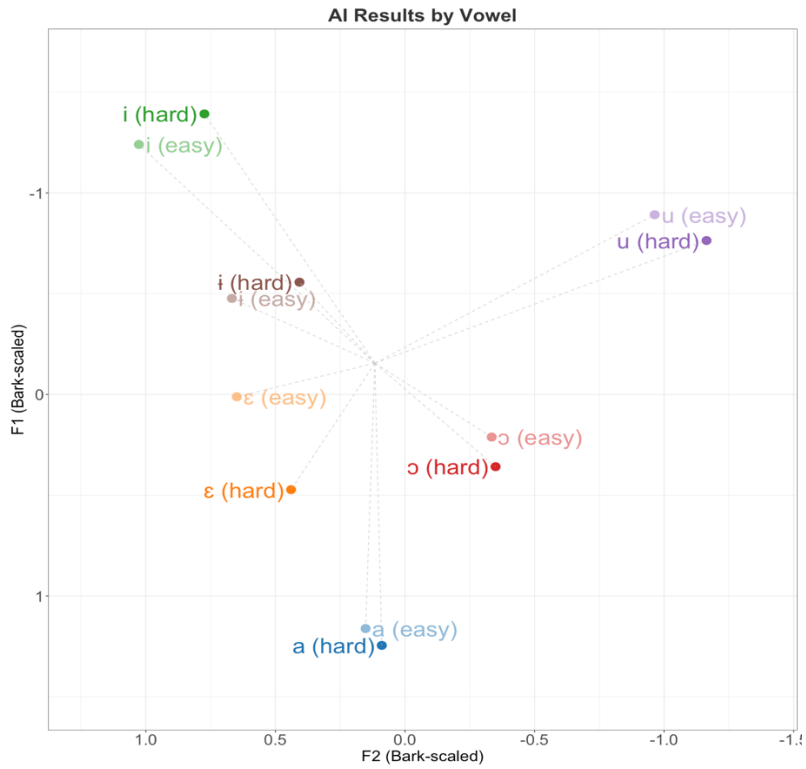

Figure 4.2. AI results in vowel space.

The LME model of the aggregated AI data revealed a significant main effect on Euclidean distance (β = .087, 95% CI [.015, .158], t(611) = 2.386, p = .017). However, this effect is considerably weaker than that in the human results (t(3016) = 13.02, p < .001). In the pairwise comparisons, /a/ and /i/ are the only two vowels that survive the Bonferroni correction, and the overall effect is likely driven by them. No effect was found for /ɛ/, /ɔ/, /u/, and /ɨ/, as shown in Table 4.4. Unlike in the human data, in which all six vowels reached statistical significance (though not uniformly), this suggests that any effect in the aggregated AI data is at best modest or inconsistent.

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	.182	2.929	105.35	< .05*
/ɛ/	.006	.058	91.13	.954
/i/	.283	2.854	47.97	< .05*
/ɔ/	.051	.712	79.8	.478
/u/	-.048	-.441	129.51	.660
/ɨ/	.126	1.311	66.66	.194

Table 4.4. AI results: per-vowel Welch t-tests comparing hard and easy words on Euclidean distance from the vowel-space centroid, expressed in Bark. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

The aggregated AI inter-vowel and intra-vowel results are reported in Tables 4.5 and 4.6, respectively. Although the LME model revealed a significant main effect (β = .071, $t(23) = 2.248$, $p < .05$*) for the inter-vowel data, no vowel survived Bonferroni correction, and the Welch t-test was non-significant ($p = .483$). The effect was considerably smaller than in the human data. No main effect was found on intra-vowel data (β = .071, t(23) =.599, $p = .555$), except for /u/, which reached Bonferroni-corrected significance, but in the opposite direction ($MD = −.560$, $t = −4.877$, $p < .05$*) (see Table 4.6).

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	.238	3.383	5.99	.015
/ɛ/	-.046	-.35	5.31	.740
/i/	.23	1.535	5.96	.176
/ɔ/	-.029	-1.428	3.61	.234
/u/	-.066	-.444	5.21	.675
/ɨ/	.097	.891	5.99	.407

Table 4.5. AI results: per-vowel Welch t-tests comparing hard and easy words in inter-vowel distance, expressed in Bark. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	.293	1.341	5.89	.229
/ɛ/	-.264	-.738	5.99	.488
/i/	.758	2.53	3.27	.079
/ɔ/	-.159	-1.453	5.78	.198
/u/	-.560	-4.877	4.00	< .05*
/ɨ/	.358	.981	4.75	.374

Table 4.6. AI results: per-vowel Welch t-tests comparing hard and easy words in intra-vowel space, expressed in Bark. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

4.5. AI Speech Results by Model

Compared with human speakers, despite a marginal statistically significant overall effect, the individual Polly and ElevenLabs voices showed little vowel-space expansion in the LME model, including a reversal of the expected PND/WF effect. The vowel /i/ for ElevenLabs hard words was notably more backed (Figure 4.4) and slightly raised in comparison to its easy vowel. In contrast, Polly's /i/ (Figure 4.3) appeared, as expected, close to its counterpart. The rest of the vowels showed negligible or contradictory effects. For example, Polly and ElevenLabs differed significantly in their /ɨ/. The /ɨ/ displayed the opposite of the expected pattern in ElevenLabs (Figure 4.4), contracting for hard words. Whereas Polly's /ɨ/ appeared more centralized (Figure 4.3), closer to the canonical /ɨ/, the /ɨ/ of ElevenLabs was more fronted, with its easy and hard variants more dispersed.

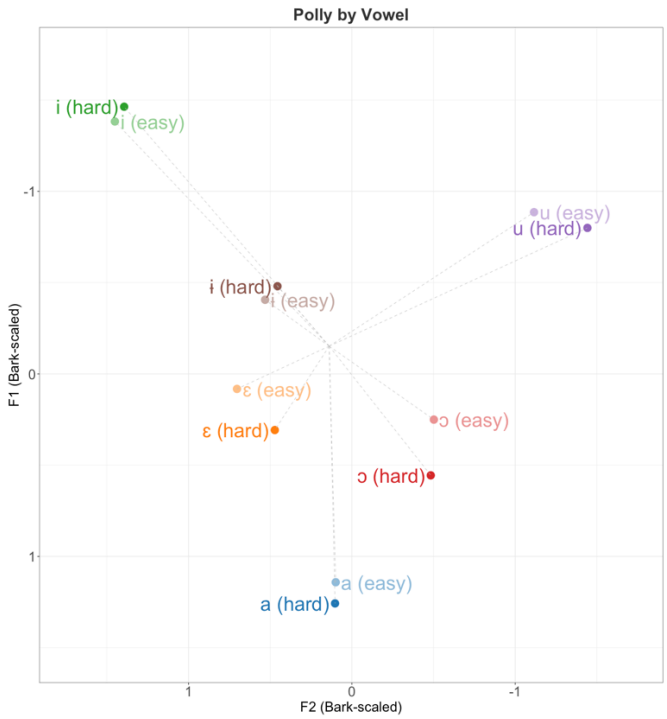

Figure 4.3. Polly: Euclidean Distances from the centroid.

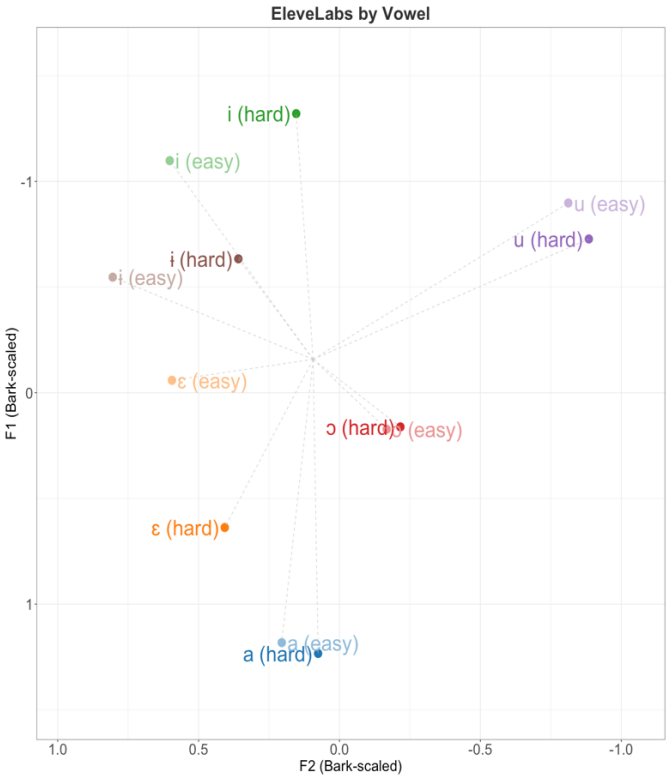

Figure 4.4. ElevenLabs: Euclidean Distances from the centroid.

Both Polly and ElevenLabs failed to reach statistical significance in the overall effect when their data was disaggregated (Polly LME: β = .083, 95% CI [−.016, .181], $t(305) = 1.658$, $p = .098$; ElevenLabs LME: β = .089, 95% CI [−.014, .192], $t(305) = 1.698$, $p = .090$). Additionally, no vowel survived Bonferroni correction in the per-vowel comparisons (Tables 4.7 and 4.8).

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	.177	2.229	62.44	.029
/ɛ/	-.225	-2.031	28.35	.052
/i/	.321	2.066	21.77	.051
/ɔ/	.158	1.862	44.67	.069
/u/	.045	.324	60.23	.747
/ɨ/	.127	1.502	25.69	.145

Table 4.7. Polly: per-vowel Welch t-tests comparing hard and easy words in inter-vowel distance, expressed in Bark. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	.185	1.977	44.01	.054
/ɛ/	.232	1.483	38.51	.146
/i/	.242	2.081	23.22	.049
/ɔ/	-.059	-.627	43.27	.534
/u/	-.144	-.961	63.73	.340
/ɨ/	.124	1.039	33.4	.306

Table 4.8. ElevenLabs: per-vowel Welch t-tests comparing hard and easy words in intra-vowel distance, expressed in Bark. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

The per-vowel data for each model's inter-vowel and intra-vowel distances are reported below. However, because there were only two voices per AI model, which resulted in a low number of tokens for the inter- and intra-vowel calculations, the lack of sufficient observations creates a very low number of degrees of freedom. Future research will provide more data to makes a meaningful comparison.

In terms of inter-vowel distances, neither Polly (β = 0.082, 95% CI [−0.023, 0.187], t(11) = 1.715, p = .114) nor ElevenLabs (β = 0.060, 95% CI [−0.035, 0.154], $t(11) = 1.391$, p = .192) reached statistical significance. Only Polly produced a vowel effect of statistical significance for /ɛ/ ($p < .05$). ElevenLabs also revealed no statistical significance in its data for inter-vowel

distance. These data are displayed in Tables 4.9 and 4.10 for Polly and ElevenLabs, respectively. Likewise, neither Polly (β = −0.040, 95% CI [−0.363, 0.284], *t(*11) = −0.270, *p* = .792) nor ElevenLabs (β = 0.182, 95% CI [−0.223, 0.587], *t(*11) = 0.988,*p*= .344) reached statistical significance for intra-vowel distance, with the exception of /u/ in the aggregated AI data, which reached Bonferroni-corrected significance (*p* = .0082) in the reversed direction, as shown in Tables 4.11 and 4.12.

Vowel	Mean Diff (Hard− Easy)	t-statistic	df	p-value
/a/	0.251	30.283	1.08	0.016
/ɛ/	-0.136	-17.603	1.99	p < 0.05 **
/i/	0.29	4.991	1.13	0.105
/ɔ/	-0.004	-0.519	1.93	0.657
/u/	-0.008	-0.123	1.89	0.914
/ɨ/	0.097	3.453	1.89	0.081

Table 4.9. Polly: Estimates for hard–easy contrasts based on inter-vowel distance in Bark, with positive values indicating greater vowel-space expansion in hard words.

Vowel	Mean Diff (Hard− Easy)	t-statistic	df	p-value
/a/	0.226	2.813	1.3	.168
/ɛ/	0.043	0.302	1.51	.799
/i/	0.17	0.902	1.23	.508
/ɔ /	-0.054	-6.036	1.06	.095
/u/	-0.123	-0.707	1.52	.572
/ɨ/	0.096	0.893	1.91	.470

Table 4.10. ElevenLabs: Estimates for hard–easy contrasts based on inter-vowel distance in Bark, with positive values indicating greater vowel-space expansion in hard words.

Vowel	Mean Diff (Hard− Easy)	t-statistic	df	p-value
/a/	0.342	1.655	1.88	.248
/ɛ/	-0.536	-5.73	1.76	.039
/i/	0.32	0.665	1.05	.622
/ɔ/	-0.172	-0.946	1.79	.454
/u/	-0.649	-2.576	1.13	.212
/ɨ/	0.456	2.034	1.00	.291

Table 4.11. Polly: Estimate for hard–easy contrasts based on intra-vowel distance in Bark, with positive values indicating greater vowel-space expansion in hard words.

Vowel	Mean Diff (Hard− Easy)	t-statistic	df	p-value
/a/	0.245	0.802	1.07	.562
/ɛ/	0.009	0.016	1.01	.990
/i/	1.196	5.629	1.14	.091
/ɔ/	-0.146	-0.763	1.31	.559
/u/	-0.471	-5.891	1.04	.101
/ɨ/	0.259	0.381	1.00	.768

Table 4.12. ElevenLabs: Estimate for hard–easy contrasts based on intra-vowel distance in Bark, with positive values indicating greater vowel-space expansion in hard words.

4.6. Radial Deviation from the Centroid in Human and Synthetic Speech

In addition to the Euclidean distance from the centroid, there is another metric to be examined regarding the degree of the hypo- and hyperarticulation effect. If hyperarticulation of vowels in hard words occurs as a speaker's method of exaggerating the vowel, it could be expected that the hyperarticulated vowel would lie on approximately the same vector from the centroid as its easy word counterpart. By taking the angle between the vector of the vowel from the hard word and that from the easy word, any significant deviation of this angle would also indicate a directional shift from the vowel target. Using the same algorithm to find the vowel points in the Euclidean distance measurements above, the angle of deviation between hard and easy words was measured.

In the human dataset, the vowels generally expanded in a nearly straight line from the center, meaning that hyperarticulation was a more extreme version of the hypoarticulated vowel. This resulted in a very low angle of deviation of several degrees or less for all vowels in the human dataset, from 0.3 degrees for /a/ up to 8 degrees for /ɔ/. In contrast, the synthesized speech resulted in much greater lateral deviation between the easy and hard variants of each vowel. This resulted in a wide range of angles from 1 degree for /a/ up to 44 degrees for /i/. Figure 4.5 illustrates the difference in angles between synthesized and human speech.

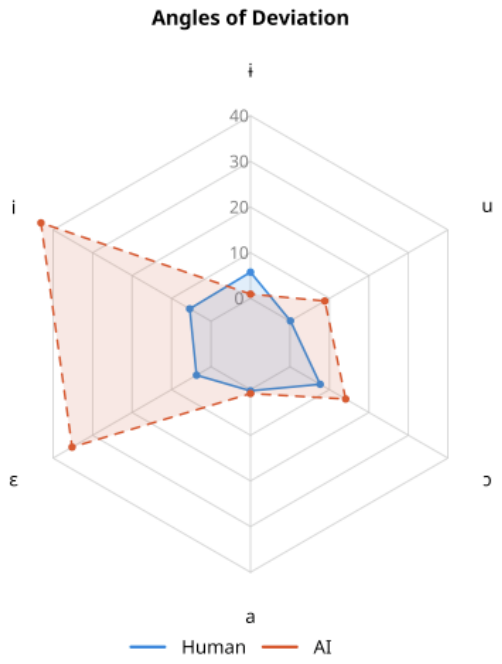

Figure 4.5. Angle of deviation between easy and hard vowels by modality.

4.7. AI Speech Recognition Activation

Research Question 3 (AI Speech Recognition): In Polish, do the internal representations of words within a transformer-based speech recognition model reflect a structure analogous to human vowel space? In particular, do 'hard' and 'easy' words occupy distinct regions within the model's representational space, corresponding with their acoustic differentiation in human speech, and do confidence metrics vary by difficulty condition like human perception?

The PCA analysis revealed a strong resemblance to the traditional vowel chart in acoustic analysis, replicating similar effects in tom Dieck et al. (2022). All vowels occupy approximately the same positions as humans in this reduction to two dimensions. In Figure 4.6, the averaged distributions of each vowel pair reveal easy and hard counterparts that are close to each other in PCA space, with some hard vowels showing outward expansion, remarkably similar to the acoustic analyses of Polish speakers above. A more nuanced view of the vowel distributions in Figure 4.7 illustrates the distribution of individual tokens.

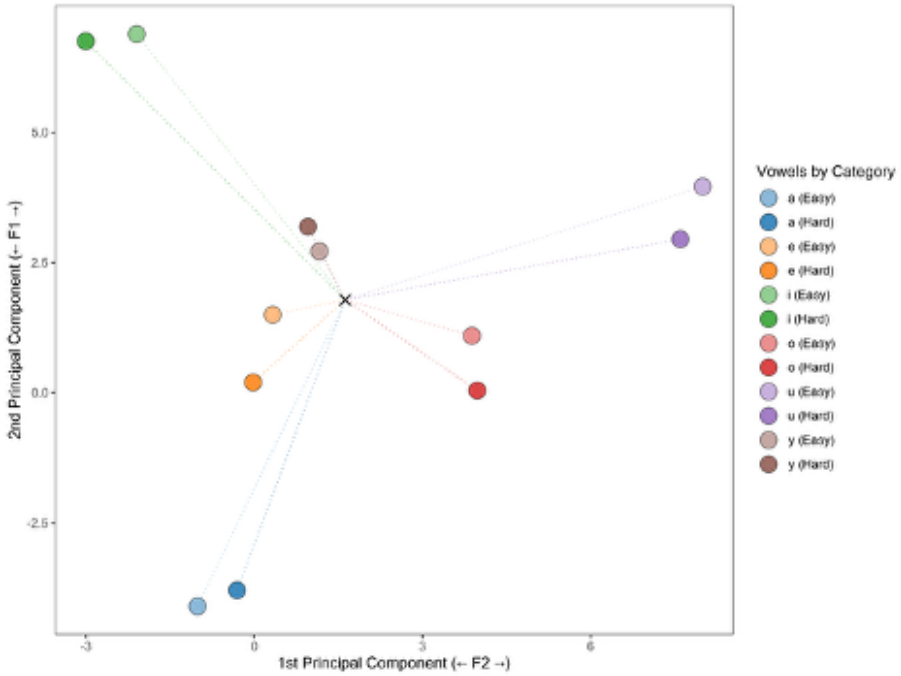

Figure 4.6. Vowels averaged by difficulty wav2vec 2.0 activations from human data.

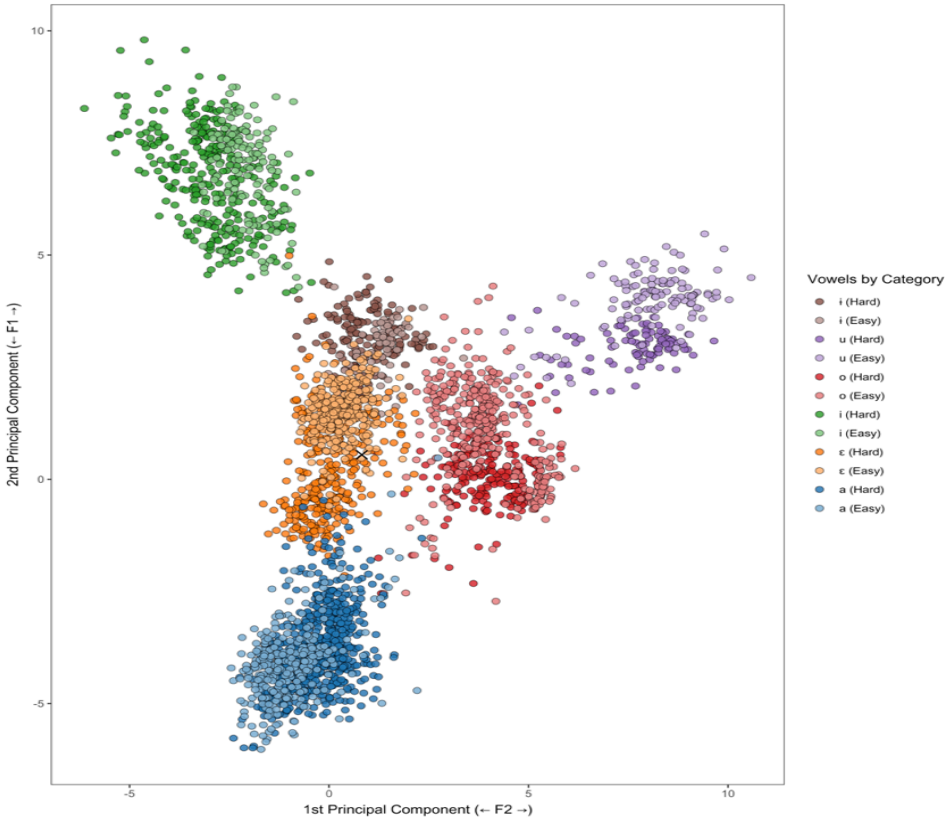

Figure 4.7. Individual wav2vec 2.vowel token activations from human data.

A Welch t-test identified a significant difference in the overall effect of expansion from the center for easy words (*MD*= -.35, 95% CI [-.51, -.19], *t(*2717.07) = -4.19, $p < .001$). The LME model confirmed this effect (β = −0.087, 95% CI [−0.154, −0.021], *t(*3042.26) = −2.58, $p < .001$). While this indicates that the vowel space expands overall in easy words, there are some per-vowel nuances that help explain this tendency.

In the pairwise comparisons, the vowels /ɛ/ (*MD* = .21, $p < .001$), /i/ (*MD* = .32, $p < .001$), and /ɨ/ (*MD* = .47, $p < .001$) in hard words all expanded outward from the PCA-based centroid with statistical significance after the Bonferroni correction. Interestingly, although/a/ (*MD* = -.53, $p < .001$) and /u/ (*MD* = -.758, $p < .001$) were statistically significant, their hard variants contracted toward the centroid, which appears to play a role in the overall effect. Only the vowel /ɔ/ (*MD*= -.04, $p = .54$) did not reach statistical significance.

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	-.528	-8.245	558.81	< .001**
/ɛ/	.207	4.847	584.16	< .001**
/i/	.324	2.849	348.08	< .05*
/ɔ/	-.042	-.615	458.07	.54
/u/	-.758	-4.796	213.77	< .001**
/ɨ/	.469	5.388	165.86	< .001**

Table 4.13. Human: per-vowel Welch t-tests comparing hard and easy words in Euclidean distance from vowel-space centroid, recreated through PCA. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

In the second set of measurements to assess how lexical properties influenced output activations of vowels in wav2vec 2.0, confidence and entropy values were extracted from the model's output layer for each target vowel and compared across the easy and hard conditions.

The LME model revealed a statistically significant main effect of the network's confidence (β = .025, 95% CI [.020, .030], *t(*3289) = 10.46, $p < .001$), with hard words (*M* = .883) eliciting higher confidence than easy words (*M* = .858), indicating greater model certainty for vowels of words with an expanded vowel space. As illustrated in Table 4.14, this effect was significant for /ɛ/, /i/, /u/, and /ɨ/ (all $p < .001$), but non-significant for /ɔ/ ($p = .020$) and for /a/ ($p = .381$), and

therefore this effect was driven only by four vowels. As in the main effect, the effect on individual vowels is relatively modest, yet statistically significant, nonetheless.

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	-.003	-.877	476.61	.381
/ɛ/	.041	8.455	481.84	< .001**
/i/	.038	4.014	216.32	< .001**
/ɔ/	.010	2.326	858.62	.020
/u/	.048	5.206	157.21	< .001**
/ɨ/	.042	5.057	212.97	< .001**

Table 4.14. Human: Welch t-tests comparing confidence in vowels in the output layer. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

The LME model similarly revealed a main effect of difficulty on entropy (β = −.083, CI [−.097, −.069], $t(3134) = -11.53$, $p < .001$), with easy words showing higher entropy ($M = .574$) than hard words ($M = .477$). As lower entropy corresponds to greater confidence, this pattern corroborates the confidence results. Both values indicate that wav2vec 2.0 processed vowels in hard words with greater certainty, as shown in Table 4.15. The vowels /ɛ/, /i/, /u/, and /ɨ/ (all $p < .001$) all achieved statistical significance, as well as /ɔ/ ($p < .05$), with no significant difference for /a/ ($p = .244$).

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	-.014	-1.166	445.04	.244
/ɛ/	-.138	-7.683	505.99	< .001**
/i/	-.118	-3.86	222.67	< .001**
/ɔ/	-.041	-2.727	876.29	< .05*
/u/	-.193	-5.863	160.26	< .001**
/ɨ/	-.15	-4.876	220.99	< .001**

Table 4.15. Human: Welch t-tests comparing entropy in vowels in the output layer. A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

Taken together, the two sets of data from the wav2vec 2.0 analysis confirmed that the model's confidence was higher for vowels of hard words. This appears counterintuitive with respect to the predicted hard-easy effects, as hard words are known to be recognized less accurately

and more slowly in humans (Luce & Pisoni, 1998). A key aspect of this result is that the confidence and entropy values reflect vowel perception, not the whole word. Therefore, one possible explanation for this effect is that a hyperarticulated vowel, which typically occurs in hard words, could produce more confidence and less entropy than its hypoarticulated counterpart in an easy word. In other words, the higher confidence may stem from the more articulated vowel, aiding in the vowel recognition, while simultaneously, there may be lower confidence at the word level. This interpretation warrants further testing in future research by comparing word-level confidence with vowel-level confidence on the same stimuli.

4.8. Conclusion

Across the studied acoustic measures, human speakers demonstrated consistent hyper- and hypoarticulation tendencies under lexical difficulty in Polish. The Euclidean distance for native speakers from the centroid was markedly greater for hard words than easy across all six vowels. Inter-vowel distance indicated system-wide expansion of vowel space in hard words, while intra-vowel distance showed no reliable effect of lexical difficulty. These results align with Stephenson's (2004) study on English. The wav2vec 2.0 speech recognition model elicited higher confidence (with corresponding lower entropy) for vowels in hard words, suggesting that the network is more sensitive to hyperarticulated vowel productions.

The radial deviation analysis further distinguished human from AI speech production. In the human data, hard and easy vowel variants expanded along virtually the same vector from the centroid, confirming that hyperarticulation represents a more extreme version of the same phonetic target rather than a categorically different one. The AI models, by contrast, showed substantially greater angular deviation between hard and easy variants — up to 44 degrees for /i/ — indicating that their hard and easy productions do not occupy the same region of vowel space, and therefore cannot be characterized as systematic expansion or contraction of a shared target.

The AI systems failed to replicate the human pattern. Neither Amazon Polly nor ElevenLabs produced the strong effects observed in the human data, and no pairwise vowel comparisons survived Bonferroni correction for either system individually (with the caveat that more data is needed for the inter- and intra-vowel comparisons). While Polish human speech exhibits consistent sensitivity to lexical difficulty at both the individual-vowel and system-wide

levels, current neural speech synthesis models fail to fully reproduce these effects. Whether this reflects a limitation of the training data or the absence of a communicative imperative is a question that the following chapter addresses in detail.

Chapter 5: Discussion

Chapter 4 presented the results of three complementary experiments designed to test the acoustic consequences of lexical difficulty in Polish. The results indicate that systematic articulatory variability (SAV) is evident in Polish speech production, but is not consistently reproduced by current AI speech synthesis models. Additionally, the internal confidence metrics of Meta's wav2vec 2.0 speech recognition model were found to be sensitive to SAV in the vowels of human speech data. Together, these results address a set of questions at the intersection of acoustic phonetics, psycholinguistics, and AI.

5.1. SAV in Polish Native Speaker Production

5.1.1. Lexical Competition and Neighborhood Effects

The acoustic analysis of Polish native speaker vowel production using Euclidean distance to the centroid supported Lindblom's (1990) theory of hyper- and hypoarticulation effects (H&H). This finding was consistent with similar studies on English (see Luce & Pisoni, 1998; Vitevitch & Luce, 1998). Vowels of hard words exhibited systematic outward expansion from the centroid, while those of easy words showed relative contraction. This finding adds Polish to the set of languages in which H&H effects have been documented.

The pairwise analysis revealed that while the H&H effect is not uniform, it is statistically significant for oral vowels in Polish. The degree of the effect with the largest mean differences observed in /ɛ/ (*MD* = .204) and /i/ (*MD* = .208), and /ɨ/ (*MD* = .085) displaying the least movement. These data suggest that Polish speakers likely hyperarticulate hard words to enhance clarity and hypoarticulate easy words to conserve articulatory effort.

Figure 5.1 Vowel space of native Polish speakers by word. Shading reflects difficulty (darker = hard, lighter = easy).

While the Welch t-test and the linear mixed effects (LME) model showed highly significant effects for the hard-easy contrast, there were a few noteworthy peculiarities. For instance, the tokens *maj* and especially *jeść* are higher and further front in vowel space compared to their respective vowel categories (see Figure5.1). Similarly, *spójrz* occupied the most peripheral fronted position within the /u/ category. These deviations are best explained by the coarticulatory encroachment of the adjacent palatal glide /j/ that creates a fronting and raising effect on the vowel target toward the /i/ target. Although the measurements were taken at the temporal midpoint of the vowel, which is typically the most stable part, these values also capture the spectral transitions in glide properties in Polish /je/, /aj/, and /uj/ sequences. While the high-front displacement of *jeść* and *spójrz* was pronounced, their underlying directionality remained consistent with the H&H theory. The vowels in *jeść* and *spójrz* were fronted and raised, though the effect of SAV was still present. *Maj* was the exception among the tokens containing a glide. Despite its hard-word

classification according to its PND- and WF-metrics, its vowel space appeared more contracted as an easy word.

Another notable exception emerged in the effects on *głos* due to its glide /w/. Despite being an easy word, the /ɔ/ in *głos* was measured closer to the periphery of the vowel space, further from the centroid than other easy tokens. The semi-vowel properties of a glide allow it to be produced without the abrupt discontinuity of a consonant and therefore maintain a relatively smooth formant transition into the following vowel (Ladefoged & Johnson, 2010). The glide transition and absence of a steady state shift the measured midpoint toward /u/, causing the word to pattern acoustically with hard words.

5.1.2. Inter-vowel Distance

To complement the findings of the centroid-based analysis, this thesis employed the inter-vowel distance metric proposed by Stephenson (2004), which measures the overall dispersion of hard and easy vowels. This vowel-space dispersion was calculated as the mean Euclidean distance between token pairs. For each difficulty set (hard or easy), every token was paired with every token of a different vowel category in the same set, and the resulting distances were averaged into a single dispersion value per condition. Though centroid-based distances are measured from a single global mean as a centroid, inter-vowel distance measures the acoustic separation among tokens throughout the vowel system — a metric more sensitive to system-wide contrast and overlap.

Corroborating Stephenson (2004) on English, the Polish results in this study showed that hard words elicited mostly greater inter-vowel dispersion than easy words. The Welch and LME models confirmed that inter-vowel distances were significantly larger in the hard condition, with the exception of /ɔ/ (see Table 5.1), which was affected by an outlier with a syllable-initial glide. The inclusion of *głos* may have altered the LME results, in that the hard-easy contrast maintained a statistically significant effect for all vowels except /ɔ/.

Vowel	Mean Difference (Hard − Easy)	*t*	*df*	*p*
/a/	.251	11.102	46.29	< .001**
/ɛ/	.084	3.696	40.43	< .0006**
/i/	.27	5.174	46.24	< .001**
/ɔ/	.064	1.855	44.34	.070
/u/	.145	4.328	46.98	< .0001**
/ɨ/	.104	3.324	47.89	< .0017**

Table 5.1. Human: per-vowel Welch t-tests comparing hard and easy words in inter-vowel distance, expressed in Bark (reproduced from Table 4.2 for reference). A Bonferroni correction was applied for 6 comparisons. Adjusted significance thresholds: $p < .0083$ (*), $p < .0017$ (**).

Intra-vowel distance was calculated as the mean Euclidean distance between all tokens of the same vowel and difficulty category in two-dimensional F1–F2 Bark space. Each token was compared to every other token within that category, and these distances were averaged to yield a single measure of in-category dispersion. This in-category variability reflects how tightly tokens are clustered, rather than their dispersion.

Consistent with Stephenson's (2004) findings on English, the current data revealed no statistically significant differences in the clustering of results in Polish. This was confirmed across both Welch ($p = .570$) and LME models ($p = .415$), suggesting that individual tokens within their difficulty category are not dispersed in a given vowel, but cluster by difficulty. In conjunction with the inter-vowel findings reported above, this pattern suggests a system-wide shift in which entire vowel categories move further from the centroid in hard words, without the tokens within each category becoming more dispersed.

5.2. SAV in AI Speech Synthesis

These results point to a gap between natural and synthetic speech in vowel production. Human vowel articulation is sensitive to conversational context: speakers shift formant targets and reduce vowels depending on factors such as speaking rate, communicative intent, and the phonetic and lexical properties of words. In contrast, modern speech synthesis models tend to produce vowels that are either overly canonical or lack the systematic variation observed in human speech (Ren et al., 2022).

5.2.1. AI-Generated Lexical Competition and Neighborhood Effects

Analysis of AI-generated speech data showed that synthetic speech in Polish failed to exhibit consistent vowel-space expansion/contraction in response to lexical difficulty. The results show that AI-generated speech, at least that tested here, lacks the SAV that is evident in human speech.

When analyzing why synthetic speech lacks SAV, it is vital to return to the fact that human speech is context dependent. Any acoustic realization of a given word is determined by a variety of factors, including how many phonologically similar competitors it has, its frequency in the lexicon, and the demands of the conversation between the speaker and listener. The effects of SAV, which are well documented in larger languages, are now supported in Polish by the current study. Despite the remarkable advances of synthetic speech models — to the point where artificial speech can be difficult to distinguish from human recordings — current models nonetheless markedly fail to exhibit the subtle spectral enhancements associated with SAV.

5.2.2. Vowel-specific and Word-specific Patterns

Synthetic speech may achieve surface-level phonetic realism, but this does not necessarily indicate cognitive modeling of speech production. Neither Amazon's nor ElevenLabs' models reproduced the expected patterns of centroid-based expansion or inter-vowel expansion, both of which were significant in the human data. This difference between AI and human speech patterns is consistent with Song et al. (2025).

If the data are considered at the token level, the differences between human and AI results become especially apparent. While human results for each vowel are closely grouped (Figure 5.4), the AI vowels are more dispersed within the same vowel category (Figure 5.5). This greater dispersion in AI production may, in theory, contribute to the perceptual difference that listeners detect as synthetic. Despite being intelligible, the variability across tokens may play a role in the perception of synthetic speech.

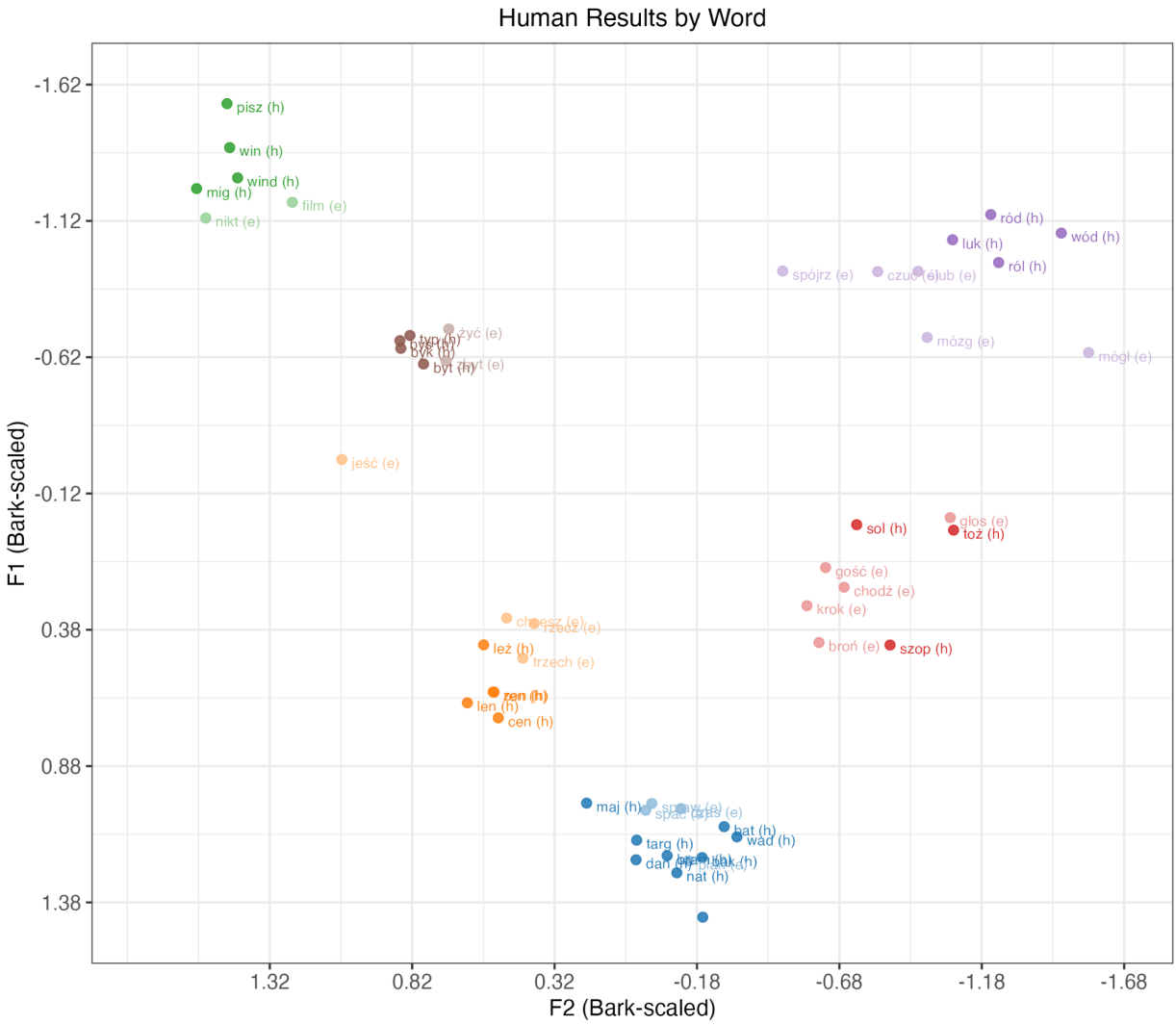

Figure 5.4. Human tokens averaged by word. Shading reflects difficulty (darker = hard, lighter = easy).

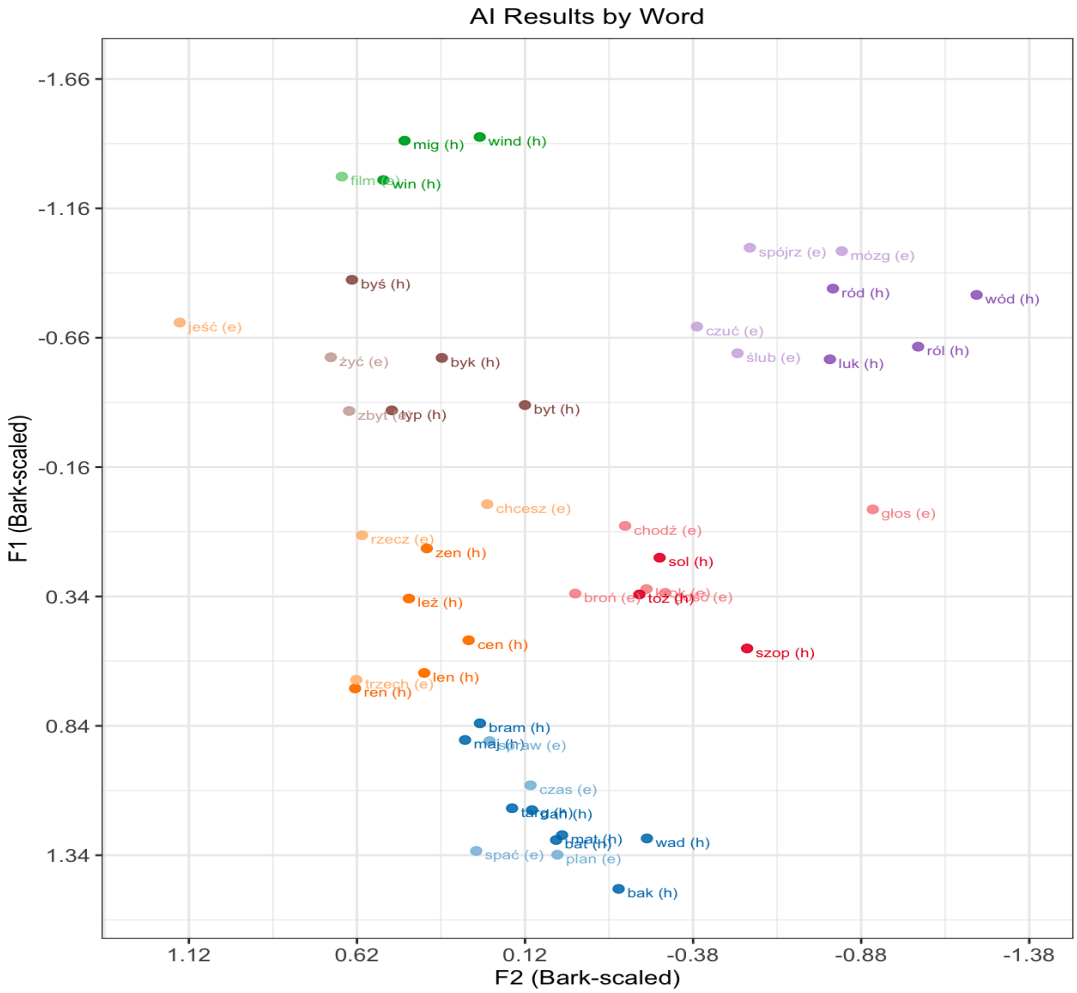

Figure 5.5. AI tokens averaged by word. Shading reflects difficulty (darker = hard, lighter = easy).

5.2.3. Inter- and Intra-vowel Effects

Similarly, p-values for LME models for inter-vowel effects were not statistically significant, with the exception of Polly's vowel /ɛ/ which reached Bonferroni-corrected significance in the opposite direction. That is, the easy /ɛ/ had a significantly more expanded inter-vowel space than the hard /ɛ/ ($t(1.99) = -17.603$). The randomness of the productions of both Amazon's and ElevenLabs' models in terms of hard/easy vowel distance from the centroid supports the conclusion that this element of human speech is not yet being produced by AI speech models on any meaningful scale.

5.2.4. Radial Effects

In addition to the Euclidean distance measurements, another interesting difference between human and AI data, which was not predicted or hypothesized in the experiment, was the deviation of the outward vectors from the centroid. If hyperarticulation of vowels in hard words occurs as a method by which speakers exaggerate the vowel, it would be expected that the hyperarticulated version of the vowel would be an outward expansion along the same approximate radial as its easy word counterpart. Calculating the angle of deviation between the radial of the vowel in the hard word and that of the easy word reveals the extent to which a vowel is, or is not, a hyper- and hypoarticulated version of its corollary. H&H theory does not predict any significant deviation in the alignment between versions of a vowel in relation to the centroid.

In the human dataset, the vowels generally expanded in a straight line from the center, meaning that hyper-articulation did appear to present a more pronounced or extreme version of the hypoarticulated vowel. Interestingly, this result did not hold true for AI speech, which saw much greater lateral deviation between hyper-and hypoarticulation, as shown in the previous chapter I Figure 4.6. The lowest level of deviation for both Polly and ElevenLabs appears in the /a, ɨ/ vowels, while the largest differences were seen in Polly's and ElevenLabs' /ɛ/ and /i/ vowels.

Previous studies have not explicitly mentioned this phenomenon (e.g., Song et al., 2025), which may be because, in human data, these vowel variants align closely along the same radial. The paucity of research on AI speech synthesis data accounts for the previously undiscovered nature of this phenomenon. The methodology of the current study enables a direct comparison of human and AI data, revealing a deviation in AI vowel production.

5.2.4.1. Theorizing AI's Non-reproduction of the Expected Effects

The rapid rise of attention-based transformer models has raised expectations of naturalness in synthetic speech. The philosophy of "scale is all you need" (a pun on the title of Vaswani et al. (2017)) posits that superior results can be achieved with more data and computing power, rather than hard-coded domain knowledge. Sutton (2019) argues that scalable learning architectures that leverage vast compute and data have historically outperformed systems built on detailed hand-

coded rules.[2] This logic suggests that a model trained on a sufficiently large corpus of natural speech should, in theory, capture all acoustic patterns in the data, including the subtler aspects of SAV.

Because the training data for commercial systems such as Polly and ElevenLabs are based on human speech, they likely contain the acoustic patterns of SAV — although the degree depends on whether training corpora capture spontaneous or only read speech. Why these effects fail to emerge in the models' outputs cannot be answered definitively without access to the models' code, weights, and training methodology, which are proprietary. Several plausible causes can nonetheless be hypothesized.

5.2.4.2. Training Data Limitations and Experiential Asymmetry

The random variation in vowel expansion and contraction observed in the AI-generated data can be explained by the fact that the training data available to Amazon and ElevenLabs is unlikely to reflect the experiential conditions that give rise to SAV in human speech in the first place. Some of these effects may be too subtle for the models to detect and generalize, but the more plausible explanation is that they are dependent on factors external to the acoustic signal. These effects in human speech are the result of a lifetime of linguistic experience, plus the totality of the environment and context in which the utterance is made. Training corpora, however large, capture only an irregular sample of this and could not conceivably record all salient situational factors for each utterance. Even if the statistical distribution of word frequency in the corpora roughly approximates natural frequencies, the decontextualization of those words may prevent the model from learning normal articulatory cues that a human does. The optimization objectives of current acoustic models may erase or obscure emergent SAV effects in the training data. Deterministic speech synthesis models tend to produce averaged outputs, which, in the case of prosody, yield flatter, less varied pitch than that produced by human speakers (Ren et al., 2022). Similar averaging in neural models, if it occurs, likely affects spectral features that are reported on in this study.

[2] One important nuance of Sutton's argument concerns the overall approach to AI — that scale tends to win over engineering — rather than the notion that scale and data capture every relevant pattern in every domain, such as speech.

5.2.4.3. Theory of Mind

Another possible inference from AI's failure to consistently reproduce SAV is that this requires something analogous to a speaker's communicative intent and understanding of the listener's needs. Lindblom's H&H theory (1990) proposes that speakers continuously adjust articulatory effort along a continuum, hyperarticulating when communicative demands are high, and hypoarticulating when they are low. The Neighborhood Activation Model (NAM) (Luce & Pisoni, 1998) provides a complementary account of why PND matters. Words with many similar-sounding competitors activate more competing lexical candidates in the listener's mind, and speakers compensate by producing these words with greater clarity. One explanation is that AI still lacks a "Theory of Mind" (Song et al., 2025), and therefore, the models do not yet appear to implement SAV in a meaningful way.

To the extent that can be inferred from publicly available information, current AI models have no explicit representation of a listener, no internal model of lexical competition, and no simulation of the vocal tract to balance articulatory economy with communicative demands. Even when the input to the speech synthesis models includes more pragmatic information,[3] awareness of the interlocutor and articulatory effort remains lacking.

Embodied cognition offers one way to understand this shortcoming. Clark (1997) argues that human cognition — including language — is distributed across brain, body, and environment. This implies that SAV arises from the interaction of cognitive factors, vocal-tract constraints, and real-time feedback during speech. Polly and ElevenLabs do not have access to these, and they map text to acoustics purely as an internal computation. Whether future models can capture SAV-like effects without an embodied component is open to empirical test.

5.2.5. Theorizing the Radial Effect

5.2.5.1. Randomization

The absence of a radial effect in AI-generated speech likely reflects foundational differences between human articulation and generative speech synthesis. Although the specific sampling mechanisms used by commercial systems such as Polly and ElevenLabs are not publicly

[3] SSML (Speech Synthesis Markup Language) is a W3C standard used in speech synthesis to control prosody rate, volume, and other characteristics in text-to-speech systems (World Wide Web Consortium, 2010).

documented, their presence can reasonably be inferred from standard transformer-based models. At inference time,[4] speech synthesis models use stochastic sampling methods to introduce variation into the output — for example, sampling-based decoding in codec language models. If either of the models studied here employs such a method, this variation would be random with respect to the lexical and phonological properties that drive SAV effects in human speech. The model may produce a slightly more or less peripheral vowel for any given token, but this fluctuation bears no systematic relationship to the word's PND or WF.

5.2.5.2. Architecture and Training

Because Amazon and ElevenLabs do not disclose the architectural details or training protocols of their speech synthesis models, any explanation for the observed deviations must remain theoretical. However, it is plausible that models are pretrained on multilingual corpora, in which high-resource languages such as English predominate, and subsequently fine-tuned to specific target languages (Gong et al., 2024; Amalas et al., 2024). If the Polish model examined here is a variant of such a base model, cross-lingual transfer may introduce instability into Polish speech productions, given that phonetic distance between pretraining and target languages has been shown to negatively affect language adaptation (Gong et al., 2024). Given the differences between English and Polish vowel inventories, this provides a plausible account for the observed angular deviations, which may reflect source-language interference rather than any systematic pattern of SAV in the Polish training data.

Rather than producing outward expansion along normative radial trajectories — as observed in human hyperarticulation — the model's variation may instead follow some of the statistical patterns inherited from the source-language training data, producing the angular deviations observed here.

[4] Inference time refers to the step in which a trained neural network model processes input and generates output, which is distinguished from the training phase. In transformer-based models, inference involves a pass through the network for each generated token.

5.3. SAV in Speech Recognition

5.3.1. Vowel Space Representation in Transformer Layers

The current study adopted the extraction methodology of tom Dieck et al. (2022). The 768-dimensional hidden activations were extracted from the last transformer layer (Layer 24) of the Polish-trained wav2vec 2.0 model, and PCA was applied to visualize the vowel space. Tom Dieck et al. focused on the spatial organization of phonetic features in the hidden layers, providing evidence that wav2vec 2.0 learns phonetic structure without explicit supervision. Extending tom Dieck et al.'s method to Polish is theoretically motivated by the hypothesis that, if wav2vec 2.0 learns universal phonetic structure as tom Dieck et al. finds, then the model should organize Polish vowels in a reconstructed vowel space to represent the effects of PND and WF. The model's internal geometry was assessed to determine how well the fundamental acoustic relationships of the F1–F2 vowel space are preserved in the embedding space of a transformer layer. Replicating the AI's vowel chart structure in Polish through PCA provides validation that the extraction method captures phonetically relevant information in a similar way. Figure 5.6 includes a scatter plot of the individual vowel tokens in the PCA analysis, which reveals a strong resemblance to a vowel chart in Hz or Bark.

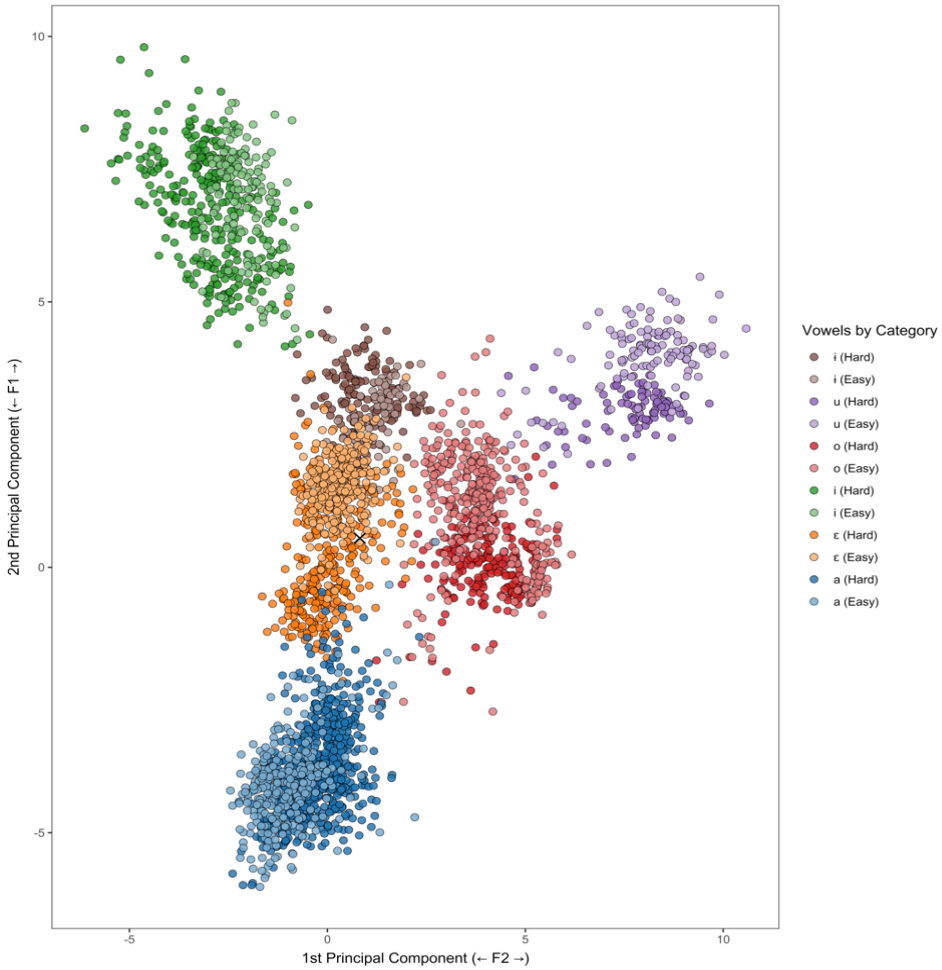

Figure 5.6. Individual wav2vec 2.0 activations from human data.

Indeed, the easy/hard categories that were present in the human data also appear in the PCA analysis in wav2vec 2.0, as shown in Figure 5.6. As discussed in the previous chapter, the wav2vec 2.0 results demonstrated statistical significance in both the Welch t-test and LME. However, this should not be construed as implying that wav2vec 2.0 contains an exact internal model of human phonetic perception or that the network faithfully simulates human cognitive processes. The vowel chart is extracted from a high-dimensional space via PCA, based on internal activations that themselves emerge from the data. Therefore, the resulting structure is best interpreted as a compressed and imperfect abstraction of how wav2vec 2.0 encodes the vowels at a particular layer in the transformer.

5.3.2. Confidence and Entropy Metrics in the Output Layer

The present study also replicated the methodology of Ravuri et al. (2024) by examining the model's confidence level at the output layer and assessing whether this confidence correlates with perceived psycholinguistic difficulty. Specifically, this thesis analyzed two output-layer metrics — maximum probability (a measure of confidence) and entropy (a measure of uncertainty) — across easy and hard word categories.

In this study, high maximum probability and low entropy values — which are inversely related and together indicate high confidence — yielded statistically significant correlations with hyperarticulated tokens in wav2vec 2.0. This suggests that hyperarticulation improves recognition, which can be explained by the more peripheral or exaggerated vowels in these tokens. The maximum probability and entropy metrics in the output layer resulted in the Welch and LME models being significant, with both at $p < .001$. In a pairwise comparison by vowel, all except /ɔ/ and /a/ had statistical significance in the observed relationship.

The present results mirror the pattern observed by Ravuri et al. (2024), who found that higher model uncertainty (i.e., higher entropy) correlates with lower-quality audio. In the current data, low entropy values indicate vowels that the model classifies as better examples of the vowel category. While humans tend to recognize hard words more slowly and less accurately than easy words due to increased lexical competition from high PND and low WF (Luce & Pisoni, 1998), the current study found that wav2vec 2.0 showed higher confidence in the vowels of hard words. This suggests that the model may be responding to the enhanced acoustic properties of an expanded vowel space, independent of lexical difficulty. This divergence is consistent with hyperarticulation research showing that expanded vowel space improves speech processing for diverse listener groups (Ferguson & Kewley-Port, 2007; Kangatharan et al., 2022). These results indicate that the model is directly sensitive to the acoustic properties — the hyperarticulated characteristics — of hard words and is less affected by PND and WF than humans.

The results may have a plausible explanation. From a purely acoustic perspective, hyperarticulated vowels are spectrally more extreme, and wav2vec 2.0 processes them with higher confidence. While wav2vec 2.0 successfully encodes phonetic structure (tom Dieck et al., 2022), its high confidence may be due to acoustic properties at the vowel level rather than the lexical attributes that cause human processing difficulty at the word level. One possible interpretation of these findings is that PND and WF effects have less impact on artificial speech recognition than

on human speech recognition. The word frequency distribution in wav2vec 2.0's training data does not necessarily match what humans are exposed to over a lifetime, which complicates a direct comparison between human perception and the model. Nevertheless, from the perspective of research on lexical and sublexical effects, the hyperarticulated vowel may have a facilitatory effect on the phonetic network, in the same way that sublexical effects can aid faster recognition when no lexical effects are present (Vitevitch & Luce, 1998). Direct comparisons with human recognition, however, require future research to study these effects more thoroughly.

5.4. Conclusion

Current speech synthesis models produce speech that is often perceptually indistinguishable from human recordings but fail to consistently reproduce SAV. However, the fact that wav2vec 2.0 exhibits higher confidence and lower entropy in the recognition of vowels of hard words, and vice versa for those of easy words, may indicate that the model is responding to hyperarticulation, which produces a more extreme vowel, rather than lexical difficulty.

Human listeners find hard words harder to recognize, despite hyperarticulated speech. While hyperarticulation may partially compensate for dense PND and sparse WF effects, it does not eliminate them. The wav2vec 2.0 model classifies the hyperarticulated vowels of hard words with higher confidence and lower entropy than the vowels of easy words, because hyperarticulated vowels are more peripheral and exaggerated, making the vowel more distinct. This greater acoustic distinctiveness of hard-word vowels yields higher model confidence (maximum probability) and lower entropy than that of easy-word vowels.

These results reveal that the gap between human and synthetic speech remains significant in terms of SAV. In speech synthesis, this gap may be explained by the fact that human speakers produce speech that is affected by both articulatory effort and listener-oriented intentions, according to the framework of H&H theory (Lindblom, 1990). However, current speech synthesis models, such as Amazon’s Polly and ElevenLabs, display very weak effects from SAV, based on the current analysis. Until speech synthesis models can more fully approximate SAV, the absence of SAV will remain one of the features that distinguish human from synthetic speech.

With regard to speech recognition, this study further supports the view that wav2vec 2.0 encodes the effects of SAV internally, in a way that somewhat resembles human perception in

terms of F1 and F2, supporting prior research (tom Dieck et al., 2022). These results suggest that speech recognition models may be responsive to SAV, whereas speech synthesis models do not yet fully replicate it in their output.

Chapter 6: Practical Applications

Speech technology has made remarkable strides in intelligibility, naturalness, and synthesis quality, yet the primary focus of dominant commercial speech systems and research remains on naturalness and human plausibility, as opposed to linguistic features such as systematic articulatory variability (SAV). Efforts are being made to refine acoustic properties such as volume, noise, and temporal structure, while key linguistic properties of speech are largely ignored. The Polish results in this study, consistent with other cross-linguistic studies, have direct, practical applications. The suggested applications in this chapter share the common premise that natural speech variability encodes information about communicative context that can be leveraged to improve speech software. Where current speech technology treats acoustic regularity as optimal, the proposals below position SAV as a novel vector for AI speech detection, bring synthetic speech closer to perceptual naturalness, improve intelligibility in recorded and accelerated media, and assist L2 learners in forming more accurate phonetic categories in the target language.

6.1. Detection of Synthetic Speech

The differences observed between human and synthetic speech in the degree of hyper- and hypoarticulation, conditioned on lexical difficulty, suggest a novel detection approach for AI-generated speech. The rapid improvement in AI speech synthesis has led to a marked increase in high-fidelity synthetic speech used for fraud, misinformation, and identity spoofing, underscoring the need for reliable automated detection methods. Voice synthesis and cloning tools are now accessible at minimal cost, and human listeners are poorly equipped to identify AI-generated speech (Barrington et al., 2025).

The growing realism of AI-generated speech has made human perception an increasingly unreliable safeguard. Barrington et al. (2025) found that listeners could correctly identify deepfake speech only 60% of the time. Listeners tend to rely on subjective impressions of naturalness — citing pauses, intonation, or breathing — suggesting no consistent or reliable detection strategy (Mai et al., 2023). As speech synthesis continues to improve, this perceptual approach will likely become even less effective, underscoring the need for improved detection methods.

Despite their growing naturalness, generative AI models still exhibit subtle, detectable artifacts in the spectral and temporal properties of speech, which automated detection systems have been developed to exploit (Jung et al., 2022; Tak et al., 2022; Yi et al., 2023). The field of automated AI speech detection has traditionally employed two approaches: feature-based and end-to-end. The first transforms raw audio into structured acoustic representations, which are then passed to a separate classifier such as a support vector machine (a supervised classification algorithm; Yi et al., 2023). More sophisticated variants also incorporate phase-based features, which capture irregular phase patterns introduced by speech synthesis, as well as prosodic features such as pitch trajectories and phoneme duration (Yi et al., 2023), which current synthesis models tend to produce with less natural variation than human speech (Ren et al., 2022). The key advantages of the feature-based approach are interpretability and computational efficiency. It requires significantly less processing power and allows the decision logic to be examined and understood, which is important in certain domains, such as forensic or legal contexts (Yang et al., 2026).

However, feature-based systems are limited in their inability to detect a wider range of irregularities produced by different speech synthesis models. The models may overfit to artifacts specific to particular synthesis methods or datasets, which can decrease their effectiveness when faced with unseen generators or real-world conditions (Liu et al., 2025).

The second approach, end-to-end models, addresses this by processing raw audio waveforms directly through a single network, allowing the model to learn which spectral patterns are most useful for AI detection without relying on human-defined rules or assumptions (Yi et al., 2023). Self-supervised learning models such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022), pre-trained on large speech corpora, are effective at learning useful representations that generalize well across datasets and languages. Combining integrate multiple feature representations and/or architectures and combine their outputs, currently represent the state of the art by compensating for the deficiencies in any single approach. Systems that combine the interpretability and computational efficiency of a feature-based approach with deep learning offer a more promising path for practical applications over single-approach systems.

Although the dataset used in this study is relatively small, and further research is needed, it provides an initial foundation for a SAV-based detection approach. Two complementary measures arise from the results of the present study, with the potential for many more linguistic

features that were out of scope. First, the degree of vowel space expansion and contraction in hard and easy words, measured as Euclidean distance from the vowel space centroid, differs systematically between human and AI-generated speech; human speakers displayed a markedly more consistent expansion for hard words and contraction for easy words than AI-generated speech. Second, the angular deviation, measured in degrees from the centroid between the vowel targets of hard and easy words, is a complementary measure that captures directional dispersion in the vowel space. Combined, the Euclidean distance and angular deviation measures from this study could constitute an effective detection method. Although formant extraction and lexical classification, which are necessary for such a method, are well-established methods and less computationally intensive than end-to-end deep learning approaches, their implementation would require further development.

This approach nevertheless has limitations. First, the AI-generated speech data in this study comprises only a sample of Amazon Polly and ElevenLabs output in Polish, which, in light of Song et al. (2025), are taken as a proxy for the broader phenomenon. The generalizability of these results across synthesis systems and languages has yet to be established. Second, the method is contingent on the presence of a sufficient number of lexically hard and easy words in the sample. If the speech to be analyzed contains fewer words than expected to be affected by lexical difficulty, the method may be unable to make a reliable prediction.

More significantly, this detection method assumes that SAV will not be successfully incorporated into future speech synthesis systems. Should synthesis systems begin to model SAV-based hyper- and hypoarticulation, this detection feature could be evaded or rendered obsolete, consistent with the existing pattern of a co-evolutionary struggle between speech synthesis and AI detection.

6.2. Enhancing the Naturalness of AI-generated Speech Synthesis

Despite achieving high naturalness in traditional ratings, more rigorous evaluations of state-of-the-art speech synthesis models still produce subtle unnatural features (Perrotin et al., 2025). Current speech synthesis models may fall short not in phonemic accuracy or vocal quality, but in the subtler acoustic properties that characterize natural human speech. At the same time, the demand for high-quality synthetic speech is growing rapidly across consumer and professional

domains where naturalness and subtlety are essential for audiences to accept a synthetic voice as a credible replacement for human performance. For example, in automated multi-lingual dubbing, AI-generated speech synthesis is increasingly deployed to replace or, at least, supplement human voice actors for large-scale media localization. Yet listeners are nonetheless sensitive to the degree of naturalness in synthesized speech. Federico et al. (2020) demonstrated that Italian listeners preferred fluent, natural-sounding speech even at the cost of reduced synchronization between dialogue and actor. The issue of reliable, intelligible, and realistic synthetic speech generation has strong implications for accessibility, including both disability rights and cross-linguistic disparities in information access (Fernández-Torné, 2016).

Compared to natural speech, synthetic speech exhibits a tendency toward temporal and spectral smoothing, due to models producing averaged rather than variable spectral outputs (Ren et al., 2022; Kögel et al., 2023). While such regularization may reduce disfluency, it also limits the production of idiosyncrasies characteristic of human speech production. Contributing to this is the fact that AI speech models are often trained on read speech, which is characterized by fewer features of spontaneous speech, such as pauses, rate variation, and varied emotion (Tan et al., 2021). As a result, this suggests that models trained on read data are insufficiently exposed to the systematic patterns that characterize conversational human speech, which likely reinforces the spectral invariance observed in systems such as Polly and ElevenLabs.

Given that the perceived unnaturalness of AI speech may stem from training data that lacks systematic lexically conditioned variation, a logical remedy is to introduce SAV by means of resynthesis. Two approaches are possible — one at training and one at inference. In the first approach during training, human speech data would be resynthesized to amplify existing SAV patterns by expanding and contracting vowel space in hard and easy words, respectively, giving the model a more explicit signal from which to learn lexically conditioned variation. In the second approach at inference, the output signal would be modified based on the PND and WF metrics of the words being produced, adjusting vowel targets post hoc. The former approach focuses on data augmentation, the latter on post-processing, but either would move speech synthesis beyond naturalness toward a more listener-oriented model of speech.

Implementing SAV in speech synthesis raises ethical concerns, particularly that more natural synthetic speech could make deepfakes harder to detect, including by the SAV-based method proposed above. At present, this prospect appears limited. The current study's findings,

together with those of Song et al. (2025), suggest that SAV is not fully implemented in commercial speech synthesis systems to date. Given the limited research on PND and WF effects in speech synthesis and recognition (e.g., Song et al., 2025), future work should examine these effects across a wider range of commercial systems before drawing broader conclusions.

6.3. SAV-Based Enhancement of Human Speech

A variety of market pressures mandate improving intelligibility in human speech without sacrificing naturalness. A growing demand for clearer dialogue in consumer media reflects both technical and behavioral shifts in audiovisual consumption. In contemporary media production, dialogue is frequently embedded in dense soundscapes where it must compete with sound effects and music — conditions that reduce intelligibility and have motivated work on automated dialogue enhancement (Paulus et al., 2019; Petermann et al., 2022). Consumer listening environments further compound this challenge. Listeners increasingly engage with media while multitasking or in noisy contexts such as public spaces and vehicles, conditions known to degrade speech intelligibility (Versfeld & Dreschler, 2002).

Another potential application arises from the widespread use of accelerated playback across video, podcast, audiobook, and e-learning platforms. For example, surveys of students who listen to recorded lectures find that they predominantly adjust the playback speed higher, with 85% of undergraduates reporting watching at faster-than-normal speeds (Murphy et al., 2022), most frequently selecting rates between 1.25x and 1.5x (Tharumalingam et al., 2025). Research suggests that acceleration up to approximately 1.5x has minimal impact on comprehension and retention, but comprehension declines markedly at 2x speed and above (Tharumalingam et al., 2025). Enhancing intelligibility at higher playback speeds could therefore extend the usable range of accelerated playback without sacrificing comprehension.

Companies have responded by expanding rate-increase options, adding fine-grained controls and implementing pitch-correction algorithms for higher speeds. Most commercial tools enhance speech by reducing background noise and reverberation, isolating dialogue, normalizing volume, and removing or shortening pauses. These approaches improve intelligibility but do not address the segmental clarity of individual speech sounds, which becomes more important for perception as speech is accelerated (Janse, 2004).

Details of the specific algorithms in use by content providers are rarely disclosed. What can be deduced from available information is that mainstream products concentrate on several general methods. Speech acceleration, or time-compressed speech, involves compressing the duration of the speech signal while preserving pitch. Methods such as Pitch-Synchronous Overlap-Add (PSOLA) shorten the speech signal by removing short frames aligned with individual pitch cycles, which preserves pitch by leaving each retained pitch cycle intact. While moderate compression retains intelligibility, higher compression rates degrade it (Janse, 2004). Neural TSM, a deep learning version of TSM, improves intelligibility by dynamically adjusting the duration of individual phonemes using an attention alignment mechanism to compress or expand speech (Lee et al., 2024). This method preserves pitch and naturalness better than conventional time-scaling methods. Another approach, used for dialogue enhancement in media, involves deep learning to isolate dialogue from background sounds (e.g., music and sound effects). The isolated speech can then be remixed with increased volume to the dialogue track to improve intelligibility (Paulus et al., 2019; Petermann et al., 2022).

Therefore, current methods for speech enhancement focus on temporal manipulation, loudness normalization, and dialogue isolation — none of which focus on the spectral characteristics of vowels and consonants that support segmental clarity. Despite evidence of the importance of vowel space in word recognition and comprehensibility (Bradlow et al., 1996), commercial systems do not appear to incorporate spectral enhancements that target vowel space. The present findings point to such an approach.

Based on the body of research linking comprehension to vowel enhancement, it follows that intelligibility could be improved by modifying the speech signal through formant resynthesis, directly manipulating F1 and F2 to expand and contract vowel space. This approach would resynthesize vowels to produce slightly more peripheral targets, which are limited in natural speech by principles of articulatory effort. Theoretically, increasing phonemic contrast and hyperarticulated vowels could improve the speed and accuracy of recognition. This could be applied broadly across the lexicon, with the degree of enhancement conditioned on factors, including but not limited to WF and PND, while preserving some variation in vowel targets.

It should be noted that some previous attempts at spectral resynthesis for intelligibility enhancement have shown limited success. Krause and Braida (2009), for example, resynthesized vowels to enhance spectral tilt by increasing amplitude in the 1000–3000 Hz range — a region

critical for F2 and F3 — but found no improvement in intelligibility. Rather than manipulating spectral tilt, the proposed approach targets F1 and F2 directly to expand vowel space, mimicking human hyperarticulation. Furthermore, earlier formant resynthesis techniques introduced artifacts that limited naturalness, but advances in vowel resynthesis now allow more controlled modifications to formant trajectories. When paired with existing enhancement methods, formant-based vowel resynthesis could meaningfully improve intelligibility for media consumers at both normal and accelerated speeds.

However, in combination with the previous two applications, the implications could be troubling. If AI synthesis systems begin to implement SAV while SAV-based detection remains in use, detection becomes unreliable in both directions: SAV-enhanced AI speech may evade detection, while enhanced human speech — modified for improved intelligibility — may be flagged as synthetic. This could have deleterious impacts on the information environment and public trust, making the liar's dividend[5] even more prevalent. Once false positives become sufficiently common, authentic content is no longer assumed to be legitimate. Such asymmetry would favor bad actors, who need only cast doubt on legitimate content.

6.4. Enhancement of Language Learning Through Adaptive Vowel Resynthesis for L2 Vowel Category Formation

The systematic relationship between vowel space expansion and lexical variables has implications beyond native speech production, including a range of pedagogical applications. When L2 learners encounter difficulties in internalizing new vowel categories, formant resynthesis could be used to exaggerate vowel contrasts acoustically, which may accelerate vowel category formation in L2 learners. Rojczyk (2011) provides a useful illustration of how L2 learners sometimes adopt maladaptive cue weighting strategies that hamper vowel category formation in the target language. Polish learners struggle to acquire the English /æ/–/ʌ/ contrast, and have been

[5] The "liar's dividend" describes the phenomenon in which growing public awareness of deep-fake technology benefits wrongdoers. When the public increasingly distrusts digital media, bad actors can evade accountability by falsely dismissing authentic content as manipulated (Chesney & Citron, 2019).

found to collapse these vowels spectrally, relying instead on duration as a compensatory cue — a strategy that reflects transfer of L1 perceptual habits rather than sensitivity to the formant-based distinctions of native English speech. Rojczyk constructed a 14-stimulus continuum that independently varied spectral properties — gradating formant frequencies across seven equal steps between /æ/ and /ʌ/ — and vowel duration, stretching each step from a short (130 ms) to a long (200 ms) version. Despite these formant gradations, Polish learners proved insensitive to the spectral shifts, instead categorizing the stimuli purely on duration, consistently identifying the 200-ms vowels as /æ/ and the 130-ms vowels as /ʌ/.

If F1 and F2 were exaggerated through formant resynthesis to maximally expand the targets of /æ/ relative to /ʌ/, learners might be able to better hear the essential contrasts that distinguish the boundaries of the English /æ/ category. This manipulation would be applied irrespective of lexical factors, since the goal is to establish the vowel category itself rather than to replicate the H&H effect. Starting with this exaggerated spectral target and then progressively reducing the degree of vowel expansion in gradations toward a typical /æ/, without varying duration, may guide learners to prioritize the correct spectral cues. Manipulating /ʌ/ in parallel — pushing it toward its own peripheral target — would further maximize the contrast between the two vowels and reinforce the acoustic distinction.

While applying vowel resynthesis — expanding and contracting formant targets to produce hyper- and hypoarticulated variants — is a relatively novel approach in L2 training, spectral exaggeration as a pedagogical technique is supported in the literature. Iverson et al. (2005) showed that extreme manipulation of formant structure could shift the perceptual strategies learners used to identify new phonetic categories. In the pretraining baseline phase of the experiment, Japanese listeners relied on secondary temporal cues to distinguish English /r/ and /l/, categorizing liquids with long closures and long transitions as /r/ and those with short closures and short transitions as /l/, as opposed to the primary spectral cue, the F3 formant. This finding aligns with Rojczyk (2011) regarding L2 learners' overreliance on secondary cues, suggesting a need for a standardized pedagogical methodology that provides learners with exaggerated primary-formant cues in the absence of confounding factors. To highlight F3 as a primary spectral cue, Iverson et al. (2005) exaggerated the differences among the F2, F3, and F4 formants for maximum salience, then gradually reduced this exaggeration over training, and finally neutralized competing secondary cues. Although learners did not fully internalize F3 as the dominant cue, the training successfully

reduced their reliance on the original durational heuristics and redirected attention toward alternative spectral information. This study demonstrates that extreme formant manipulation can reweight cue use in L2 learners, even if it does not guarantee complete acquisition of the native cue hierarchy.

While a full training plan is beyond the scope of this thesis, these findings suggest a path for future development. A system incorporating vowel resynthesis and perceptual fading could provide a structured framework for L2 phonetic category training. Further research is needed to determine the optimal degree of vowel space exaggeration, the appropriate fading algorithm, and whether the approach generalizes across L2 contrasts and language pairs.

6.5. Conclusion

This chapter has argued that systematic articulatory variability (SAV) has a variety of applications. If models trained predominantly on systematically solicited speech fail to internalize lexically conditioned articulatory dynamics, then introducing SAV — both through data augmentation and through controlled resynthesis — could move synthesis models beyond acoustic fidelity toward communicative realism. At the same time, the fact that SAV is not currently evident in artificially generated speech creates an equally plausible pathway to more effective AI speech-detection software. This analysis recognizes the inherent competition between commercially available AI-based synthesis and detection systems and the ethical implications of implementing SAV in both domains.

In human speech enhancement, this chapter has argued that current commercial solutions focus on temporal compression, loudness normalization, and source separation, while leaving the phonemic structure of speech largely untouched. Given that intelligibility under acceleration is dependent on phonemic contrast, particularly in high-speed or noisy listening conditions, vowel space expansion — independent of lexical properties — offers a linguistically plausible alternative to speech enhancement.

Three of these applications create a trilateral ethical dilemma. A SAV-based AI detection feature will produce false negatives as artificial synthesis begins to model SAV. If human speech enhancement also introduces exaggerated SAV to improve intelligibility, the same acoustic signatures will create false positives in AI detection. This interaction among detection, synthesis,

and enhancement has no simple solution, as improvements in any one domain affect the performance of the others. As AI-generated speech proliferates and SAV-based products begin to enter the market, the asymmetric deployment of these tools could have a profoundly negative effect on the information environment.

Finally, the chapter considers the positive implications of SAV-based tools on accessibility and language learning. In L2 phonetic training, the chapter extended the logic of vowel space manipulation to the development of perceptual cues. Evidence that learners may overweight secondary cues, such as duration, while neglecting primary spectral cues, supports the use of exaggerated formant resynthesis as a structured training paradigm.

The applications outlined in this chapter argue for greater attention to SAV in future speech research. Speech variability has long posed a challenge for developers of speech applications (Benzeghiba et al., 2007). If SAV encodes communicative intent and lexical structure, then future speech systems — whether for synthesis, enhancement, detection, or pedagogy — must treat systematic variability as a necessary feature of speech rather than random variation. The central contribution of this work is to document the presence of SAV in Polish, the failure of current AI speech synthesis to fully reproduce it, the sensitivity of wav2vec 2.0 to acoustic correlates of SAV, and the implications of these findings for speech technology applications. In repositioning SAV from a linguistic finding to a design principle, this thesis argues for a linguistically grounded approach in speech technology — one in which human speech variability is not smoothed over by commercial models, but understood, modeled, and strategically amplified where appropriate.

References

Ada, A. A., Jørgensen, S. H., & Fritsch, J. (2024). Cultures of the AI paralinguistic in voice cloning tools. *Companion Publication of the 2024 ACM Designing Interactive Systems Conference, DIS '24 Companion*, 249–252. https://doi.org/10.1145/3656156.3663708

Ahlawat, H., Aggarwal, N., & Gupta, D. (2025). Automatic Speech Recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*, *6*, 201–237. https://doi.org/10.1016/j.ijcce.2024.12.007

Alzahrani, A. (2025). Jiwar: A database and calculator for word neighborhood measures in 40 languages. *Behavior Research Methods*, *57*(3), 98. https://doi.org/10.3758/s13428-025-02612-7

Amalas, A., Ghogho, M., Chetouani, M., & Oulad Haj Thami, R. (2024). *A multilingual training strategy for low resource text to speech*. https://arxiv.org/abs/2409.01217

Amazon Web Services. (2024). *Neural voices. Amazon Polly documentation*. https://docs.aws.amazon.com/polly/latest/dg/neural-voices.html

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common Voice: A massively-multilingual speech corpus. *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, 4218–4222.

Arutiunian, V., & Lopukhina, A. (2020). The effects of phonological neighborhood density in childhood word production and recognition in Russian are opposite to English. *Journal of Child Language*, *47*(6), 1244–1262. https://doi.org/10.1017/S0305000920000112

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2006.11477

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Barrington, S., Cooper, E. A., & Farid, H. (2025). People are poorly equipped to detect AI-powered voice clones. *Scientific Reports*, *15*(1), 11004. https://doi.org/10.1038/s41598-025-94170-3

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111. https://doi.org/10.1016/j.jml.2008.06.003

Benesty, J., Sondhi, M. M., & Huang, Y. A. (Eds.). (2008). *Springer Handbook of Speech Processing*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-49127-9

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Intrinsic Speech Variations*, *49*(10), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Berghoff, R., & Bylund, E. (2025). Diversity in research on the psychology of language: A large-scale examination of sampling bias. *Cognition*, *256*, 106043. https://doi.org/10.1016/j.cognition.2024.106043

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

BNC Consortium. (2007). *The British National Corpus, XML Edition* [Dataset]. Oxford Text Archive. http://www.natcorp.ox.ac.uk/

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer (Version 6.3.09)* [Computer software]. http://www.praat.org/

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074–2085. https://doi.org/10.1121/1.427952

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3–4), 255–272. https://doi.org/10.1016/S0167-6393(96)00063-5

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–479.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Brysbaert, M., Lagrou, E., & Steven, M. (2017). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition*, *20*(3), 530–548. https://doi.org/10.1017/S1366728916000353

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for

American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Butzberger, J., Murveit, H., Shriberg, E., & Price, P. (1992). Spontaneous speech effects in large vocabulary speech recognition applications. *Proceedings of the DARPA Speech and Natural Language Workshop*, 339–343.

Bybee, J. (2001). *Phonology and Language Use* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511612886

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022). WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war. *Foreign Affairs*, *98*(1), 147–155.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. Harper & Row.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. MIT Press.

Cohn, M., & Zellou, G. (2020). Perception of Concatenative vs. Neural Text-To-Speech (TTS): Differences in Intelligibility in Noise and Language Attitudes. *Interspeech 2020*, 1733–1737. https://doi.org/10.21437/Interspeech.2020-1336

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. *Interspeech 2021*, 2426–2430. https://doi.org/10.21437/Interspeech.2021-329

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)* [Dataset]. https://www.english-corpora.org/coca/

Dean, J. (2022). A Golden Decade of Deep Learning: Computing Systems & Applications. *Daedalus*, *151*(2), 58–74. https://doi.org/10.1162/daed_a_01900

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, *66*(5), 843–863. https://doi.org/10.1080/17470218.2012.720994

Dong, L., Xu, S., & Xu, B. (2018). Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5884–5888.

Dufour, S., & Frauenfelder, U. H. (2010). Phonological neighbourhood effects in French spoken-word recognition. *Quarterly Journal of Experimental Psychology*, *63*(2), 226–238. https://doi.org/10.1080/17470210903308336

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Speech Reduction*, *39*(3), 253–260. https://doi.org/10.1016/S0095-4470(11)00055-6

Federico, M., Enyedi, R., Barra-Chicote, R., Giri, R., Isik, U., Krishnaswamy, A., & Sawaf, H. (2020). From Speech-to-Speech Translation to Automatic Dubbing. *Proceedings of the 17th International Conference on Spoken Language Translation*, 257–264. https://doi.org/10.18653/v1/2020.iwslt-1.31

Ferguson, S., & Kewley-Port, D. (2007). Talker Differences in Clear and Conversational Speech: Acoustic Characteristics of Vowels. *Journal of Speech, Language, and Hearing Research*, *50*(5), 1241–1255. https://doi.org/10.1044/1092-4388(2007/087)

Fernández-Torné, A. (2016). *Audio description and technologies: Study on the semi-automatisation of the translation and voicing of audio descriptions* [Doctoral dissertation]. Universitat Autònoma de Barcelona.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. https://doi.org/10.1016/j.jml.2011.11.006

Gales, M., & Young, S. (2008). The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing*, *1*(3), 195–304. https://doi.org/10.1561/2000000004

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471. https://doi.org/10.1162/089976600300015015

Goldstein, A., Grinstein-Dabush, A., Schain, M., Wang, H., Hong, Z., Aubrey, B., Nastase, S. A., Zada, Z., Ham, E., Feder, A., Gazula, H., Buchnik, E., Doyle, W., Devore, S., Dugan, P., Reichart, R., Friedman, D., Brenner, M., Hassidim, A., … Hasson, U. (2024). Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, *15*(1), 2768. https://doi.org/10.1038/s41467-024-46631-y

Gong, C., Cooper, E., Wang, X., Qiang, C., Geng, M., Wells, D., Wang, L., Dang, J., Tessier, M., Pine, A., Richmond, K., & Yamagishi, J. (2024). An Initial Investigation of Language Adaptation for TTS Systems under Low-resource Scenarios. *Interspeech 2024*, 4963–4967. https://doi.org/10.21437/Interspeech.2024-969

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2018). *Semantic projection: Recovering human knowledge of multiple, distinct object features from word embeddings* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1802.01241

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent eural networks. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 369–376. https://doi.org/10.1145/1143844.1143891

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5–6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Grosman, J. (2021). *Fine-tuned XLSR-53 large model for speech recognition in Polish* [Computer software]. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-polish

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. *Interspeech 2020*, 5036–5040. https://doi.org/10.21437/Interspeech.2020-3015

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech

Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America*, *118*(5), 3267–3278. https://doi.org/10.1121/1.2062307

Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Janse, E. (2004). Word perception in fast speech: Artificially time-compressed vs. naturally produced fast speech. *Speech Communication*, *42*(2), 155–173. https://doi.org/10.1016/j.specom.2003.07.001

Jassem, W. (1992). Acoustic-phonetic variability of Polish vowels. *Archives of Acoustics*, *17*(2), 217–233.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer. https://doi.org/10.1007/b98835

Jung, J., Heo, H.-S., Tak, H., Shim, H., Chung, J. S., Lee, B.-J., Yu, H.-J., & Evans, N. (2022). AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6367–6371. https://doi.org/10.1109/ICASSP43922.2022.9747766

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

Kangatharan, J., Uther, M., & Gobet, F. (2022). The effect of hyperarticulation on speech comprehension under adverse listening conditions. *Psychological Research*, *86*(5), 1535–1546. https://doi.org/10.1007/s00426-021-01595-2

Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. =. *Speech Communication*, *27*(3), 187–207. https://doi.org/10.1016/S0167-6393(98)00085-5

Kaźmierski, K. (2019). Durational variation in Polish fricatives provides evidence for hybrid models of phonology. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1997–2001). Australasian Speech Science and Technology Association.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857. https://doi.org/10.1121/1.398894

Kögel, F., Nguyen, B., & Cardinaux, F. (2023). Towards Robust FastSpeech 2 by Modelling Residual Multimodality. *INTERSPEECH 2023*, 4309–4313. https://doi.org/10.21437/Interspeech.2023-879

Kong, J., Kim, J., & Bae, J. (2020). *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2010.05646

Krause, J. C., & Braida, L. D. (2009). Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech. *The Journal of the Acoustical Society of America*, *125*(5), 3346–3357. https://doi.org/10.1121/1.3097491

Kudela-Dobrogowska, K. (1973). Further studies of the optimal formant frequency values of Polish vowels. In W. Jassem (Ed.), *Speech Analysis and Synthesis* (Vol. 3, pp. 265–285). PWN.

Ladefoged, P., & Johnson, K. (2010). *A course in phonetics*. Cengage Learning.

Laptev, A., & Ginsburg, B. (2022). Fast entropy-based methods of word-level confidence estimation for end-to-end automatic speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)* (pp. 152–159). IEEE. https://arxiv.org/abs/2212.08703

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, J., Jang, S., & Chang, J.-H. (2024). Neural ATSM: Fully Neural Network-based Adaptive Time-Scale Modification Using Sentence-Specific Dynamic Control. *Interspeech 2024*, 4903–4907. https://doi.org/10.21437/Interspeech.2024-2380

Lee, S. J., Kang, B.-O., Chung, H., Park, J. G., & Lee, Y. K. (2018). *Hypo and Hyperarticulated Speech Data Augmentation for Spontaneous Speech Recognition*. 2018 26th European Signal Processing Conference (EUSIPCO). https://doi.org/10.23919/EUSIPCO.2018.8553555

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Springer Netherlands. https://doi.org/10.1007/978-94-009-2037-8_16

Liu, C., Xu, X., & Xiao, F. (2025). ASSD: An AI-Synthesized Speech Detection Scheme Using Whisper Feature and Types Classification. *IEEE Transactions on Audio, Speech and Language Processing*, *33*, 542–556. https://doi.org/10.1109/TASLPRO.2024.3520385

Lorenzo-Trueba, J., Drugman, T., Latorre, J., Merritt, T., Putrycz, B., Barra-Chicote, R., Moinet, A., & Aggarwal, V. (2019). Towards Achieving Robust Universal Neural Vocoding. *Proceedings of Interspeech 2019*, 181–185. https://doi.org/10.21437/Interspeech.2019-1424

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE*, *18*(8), e0285333. https://doi.org/10.1371/journal.pone.0285333

Malisz, Z., Brandt, E., Möbius, B., Oh, Y. M., & Andreeva, B. (2018). Dimensions of Segmental Variability: Interaction of Prosody and Surprisal in Six Languages. *Frontiers in Communication*, *3*, 25. https://doi.org/10.3389/fcomm.2018.00025

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2014). SUBTLEX-PL: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, *47*. https://doi.org/10.3758/s13428-014-0489-4

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE*, *7*(8), e43230. https://doi.org/10.1371/journal.pone.0043230

Markov, A. A. (1906). Extension of the law of large numbers to dependent quantities. *Izvestiia Fiziko-Matematicheskogo Obschestva Pri Kazanskom Universitete*, *2*(15), 135–156.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63. https://doi.org/10.1016/0010-0285(78)90018-X

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. *Information Fusion*, *99*, 101869. https://doi.org/10.1016/j.inffus.2023.101869

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In L. Vanderwende, H. Daumé III, & K. Kirchhoff (Eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Association for Computational Linguistics. https://aclanthology.org/N13-1090/

Mitchell, M. (2021). *Why AI is Harder Than We Think* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2104.12871

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, *38*(8), 114–117.

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, *E99.D*(7), 1877–1884. https://doi.org/10.1587/transinf.2015EDP7457

Munson, B., & Solomon, N. P. (2004). The Effect of Phonological Neighborhood Density on Vowel Articulation. *Journal of Speech, Language, and Hearing Research*, *47*(5), 1048–1058. https://doi.org/10.1044/1092-4388(2004/078)

Murphy, D. H., Hoover, K. M., Agadzhanyan, K., Kuehn, J. C., & Castel, A. D. (2022). Learning in double time: The effect of lecture video speed on immediate and delayed comprehension. *Applied Cognitive Psychology*, *36*(1), 69–82. https://doi.org/10.1002/acp.3899

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, *80*(5), 1297–1308. https://doi.org/10.1121/1.394433

Nerbonne, J., & Heeringa, W. (1997). Measuring Dialect Distance Phonetically. *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*. https://aclanthology.org/W97-1102/

Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L.-J. (2019). A Review of Deep Learning Based Speech Synthesis. *Applied Sciences*, *9*(19), 4050. https://doi.org/10.3390/app9194050

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *WaveNet: A Generative Model for Raw Audio* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1609.03499

Owens, J. D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A. E., & Purcell, T. J. (2007). A Survey of General-Purpose Computation on Graphics Hardware. *Computer Graphics Forum*, *26*(1), 80–113. https://doi.org/10.1111/j.1467-8659.2007.01012.x

Paulus, J., Torcoli, M., Uhle, C., Herre, J., Disch, S., & Fuchs, H. (2019). Source Separation for Enabling Dialogue Enhancement in Object-based Broadcast with MPEG-H. *Journal of the Audio Engineering Society*, *67*(7/8), 510–521. https://doi.org/10.17743/jaes.2019.0032

Perrotin, O., Stephenson, B., Gerber, S., Bailly, G., & King, S. (2025). Refining the evaluation of speech synthesis: A summary of the Blizzard Challenge 2023. *Computer Speech & Language*, *90*, 101747. https://doi.org/10.1016/j.csl.2024.101747

Petermann, D., Wichern, G., Wang, Z.-Q., & Roux, J. L. (2022). The Cocktail Fork Problem: Three-Stem Audio Separation for Real-World Soundtracks. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 526–530. https://doi.org/10.1109/ICASSP43922.2022.9746005

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. L. Bybee & P. J. Hopper (Eds.), *Typological Studies in Language* (Vol. 45, pp. 137–158). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.45.08pie

Port, R. F. (2010). Language as a Social Institution: Why Phonemes and Words Do Not Live in the Brain. *Ecological Psychology*, *22*(4), 304–326. https://doi.org/10.1080/10407413.2010.517122

Prenger, R., Valle, R., & Catanzaro, B. (2019). Waveglow: A Flow-based Generative Network for Speech Synthesis. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3617–3621. https://doi.org/10.1109/ICASSP.2019.8683143

Przepiórkowski, A. (Ed.). (2012). *Narodowy Korpus Języka Polskiego: Praca zbiorowa*. Wydawnictwo Naukowe PWN.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. https://doi.org/10.1109/5.18626

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, *202*, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html

Raitio, T., Rasipuram, R., & Castellani, D. (2020). Controllable Neural Text-to-Speech Synthesis Using Intuitive Prosodic Features. *Proceedings of Interspeech 2020*, 4432–4436. https://doi.org/10.21437/Interspeech.2020-2861

Ravuri, A., Cooper, E., & Yamagishi, J. (2024). Uncertainty as a predictor: Leveraging self-supervised learning for zero-shot MOS prediction. *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW): Self-Supervision in Audio, Speech and Beyond (SASB) Workshop*.

Ren, Y., Tan, X., Qin, T., Zhao, Z., & Liu, T.-Y. (2022). Revisiting Over-Smoothness in Text to Speech. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8197–8213. https://doi.org/10.18653/v1/2022.acl-long.564

Rojczyk, A. (2011). Overreliance on duration in nonnative vowel production and perception: The within lax vowel category contrast. In M. Wrembel, M. Kul, & K. Dziubalska-Kołaczyk (Eds.), *Achievements and Perspectives in SLA of Speech: New Sounds 2010* (Vol. 2, pp. 239–249). Peter Lang.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press. https://doi.org/10.7551/mitpress/5236.001.0001

Sanders, N. C., & Chin, S. B. (2009). Phonological Distance Measures. *Journal of Quantitative Linguistics*, *16*(1), 96–114. https://doi.org/10.1080/09296170802514138

Scarborough, R., Fougeron, C., & Marques, L. (2018). Neighborhood-conditioned coarticulation effects in French listener-directed speech. *The Journal of the Acoustical Society of America*, *144*(3_Supplement), 1900–1900. https://doi.org/10.1121/1.5068323

Scarborough, R., & Zellou, G. (2013). Clarity in communication: “Clear” speech authenticity and lexical neighborhood density effects in speech production and perception. *The Journal of the Acoustical Society of America*, *134*(5), 3793–3807. https://doi.org/10.1121/1.4824120

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Schwartz, G. (2010). Phonology in the Speech Signal—Unifying cue and Prosodic Licensing. *Poznań Studies in Contemporary Linguistics*, *46*(4), 499–518. https://doi.org/10.2478/v10010-010-0025-3

Schwartz, G. (2021). The phonology of vowel VISC-osity – acoustic evidence and representational implications. *Glossa: A Journal of General Linguistics*, *6*(1). https://doi.org/10.5334/gjgl.1182

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368

Siew, C. S. Q., & Vitevitch, M. S. (2019). The phonographic language network: Using network science to investigate the phonological and orthographic similarity structure of language. *Journal of Experimental Psychology: General*, *148*(3), 475–500. https://doi.org/10.1037/xge0000575

Song, J. Y., Rojas, C., & Pycha, A. (2025). Factors modulating perception and production of speech by AI tools: A test case of Amazon Alexa and Polly. *Frontiers in Psychology*, *16*, 1520111. https://doi.org/10.3389/fpsyg.2025.1520111

Stephenson, L. (2004). Lexical frequency and neighbourhood density effects on vowel production in words and nonwords. *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, 364–369. https://assta.org/

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *The Journal of the Acoustical Society of America*, *74*(3), 695–705. https://doi.org/10.1121/1.389855

Sutton, R. (2019, March 13). *The bitter lesson. Incomplete Ideas*. http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Tak, H., Todisco, M., Wang, X., Jung, J., Yamagishi, J., & Evans, N. (2022). Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *Proceedings of Odyssey*, 112–119.

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). *A Survey on Neural Speech Synthesis* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2106.15561

Tånnander, C., House, D., & Edlund, J. (2023). Multi-tracking TTS for the development and post hoc interpretation of neural TTS. In *Proceedings of the 13th International Conference of Nordic Prosody* (pp. 85–93). Sciendo. https://doi.org/10.2478/9788366675728-006

Taylor, P. (2009). *Text-to-Speech Synthesis* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511816338

Tharumalingam, T., Roberts, B. R. T., Fawcett, J. M., & Risko, E. F. (2025). Increasing Video Lecture Playback Speed Can Impair Test Performance – a Meta-Analysis. *Educational Psychology Review*, *37*(2), 35. https://doi.org/10.1007/s10648-025-10003-9

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. *Acoustics, Speech, and Signal*

Processing, IEEE International Conference On, *3*, 1315–1318. https://doi.org/10.1109/ICASSP.2000.861820

tom Dieck, T., Pérez-Toro, P. A., Arias, T., Noeth, E., & Klumpp, P. (2022). Wav2vec behind the Scenes: How end2end Models learn Phonetics. *Interspeech 2022*, 5130–5134. https://doi.org/10.21437/Interspeech.2022-10865

Tomaschek, F., Wieling, M., Arnold, D., & Baayen, R. H. (2013). Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience. *Proceedings of Interspeech*, 1302–1306.

Ton Dijkstra, Grainger, J., & van Heuven, W. J. B. (1999). Recognition of Cognates and Interlingual Homographs: The Neglected Role of Phonology. *Journal of Memory and Language*, *41*(4), 496–518. https://doi.org/10.1006/jmla.1999.2654

Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021). A Brief History of AI: How to Prevent Another Winter (A Critical Review). *PET Clinics*, *16*(4), 449–469. https://doi.org/10.1016/j.cpet.2021.07.001

Umeda, N. (1975). Vowel duration in American English. *The Journal of the Acoustical Society of America*, *58*(2), 434–445. https://doi.org/10.1121/1.380688

Ustinov, A., Yordanov, M., Kuchma, A., & Bychkov, M. (2025). *Cross-Technology Generalization in Synthesized Speech Detection: Evaluating AST Models with Modern Voice Generators*. (arXiv:2503.22503). https://arxiv.org/abs/2503.22503

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. *NIPS'17*, 6000–6010.

Versfeld, N. J., & Dreschler, W. A. (2002). The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. *The Journal of the Acoustical Society of America*, *111*(1), 401–408. https://doi.org/10.1121/1.1426376

Vitevitch, M. S., & Luce, P. A. (1998). When Words Compete: Levels of Processing in Perception of Spoken Words. *Psychological Science*, *9*(4), 325–329. https://doi.org/10.1111/1467-9280.00064

Vitevitch, M. S., & Luce, P. A. (1999). Phonotactics, Neighborhood Activation, and Lexical Access for Spoken Words. *Brain and Language*, *68*(1–2), 306–311. https://doi.org/10.1006/brln.1999.2116

Vitevitch, M. S., & Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders*, *3*(1), 64–73. https://doi.org/10.1080/14769670400027332

Vitevitch, M. S., & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes*, *21*(6), 760–770. https://doi.org/10.1080/01690960500287196

World Wide Web Consortium. (2010). *Speech Synthesis Markup Language (SSML) Version 1.1*. W3C. https://www.w3.org/TR/speech-synthesis11/

Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI* (pp. 75–87). Cambridge University Press. https://doi.org/10.1017/CBO9780511486425.005

Yang, T., Sun, C., Lyu, S., & Rose, P. (2026). Forensic deepfake audio detection using segmental speech features. *Forensic Science International*, *379*, 112768. https://doi.org/10.1016/j.forsciint.2025.112768

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Yates, M., Locker, L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, *11*(3), 452–457. https://doi.org/10.3758/BF03196594

Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023). *Audio deepfake detection: A survey*. https://arxiv.org/abs/2308.14970

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*(11), 1039–1064. https://doi.org/10.1016/j.specom.2009.04.004

Ziegler, J. C., Muneaux, M., & Grainger, J. (2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, *48*(4), 779–793. https://doi.org/10.1016/S0749-596X(03)00006-8

Appendix: Stimulus List

Word	Class	WF	PND
słów	easy	HF	LD
mózg	easy	HF	LD
krok	easy	HF	LD
czuć	easy	HF	LD
jeść	easy	HF	LD
spraw	easy	HF	LD
ślub	easy	HF	LD
głos	easy	HF	LD
gość	easy	HF	LD
spać	easy	HF	LD
trzech	easy	HF	LD
film	easy	HF	LD
plan	easy	HF	LD
spójrz	easy	HF	LD
żyć	easy	HF	LD
broń	easy	HF	LD
rzecz	easy	HF	LD
mógł	easy	HF	LD
zbyt	easy	HF	LD

Word	Class	WF	PND
nikt	easy	HF	LD
czas	easy	HF	LD
chcesz	easy	HF	LD
chodź	easy	HF	LD
byś	hard	LF	HD
typ	hard	LF	HD
win	hard	LF	HD
bram	hard	LF	HD
byk	hard	LF	HD
ród	hard	LF	HD
maj	hard	LF	HD
pisz	hard	LF	HD
wad	hard	LF	HD
ren	hard	LF	HD
bat	hard	LF	HD
targ	hard	LF	HD
nat	hard	LF	HD
sol	hard	LF	HD
zen	hard	LF	HD
cen	hard	LF	HD
ról	hard	LF	HD
leż	hard	LF	HD

Word	Class	WF	PND
len	hard	LF	HD
wód	hard	LF	HD
kół	hard	LF	HD
mat	hard	LF	HD
bak	hard	LF	HD
toż	hard	LF	HD
wind	hard	LF	HD
byt	hard	LF	HD
luk	hard	LF	HD
dań	hard	LF	HD
mig	hard	LF	HD
szop	hard	LF	HD
chrzest	distractor	LF	LD
wziął	distractor	HF	LD
zięć	distractor	LF	LD
smycz	distractor	LF	LD
gryźć	distractor	LF	LD
gęś	distractor	LF	LD
sztab	distractor	LF	LD
kłam	distractor	LF	LD
młot	distractor	LF	LD
ząb	distractor	LF	LD

Word	Class	WF	PND
wąż	distractor	LF	LD
wręcz	distractor	LF	LD
skecz	distractor	LF	LD
zniszcz	distractor	LF	LD
pięść	distractor	LF	LD
chęć	distractor	LF	LD
część	distractor	HF	LD
łyk	distractor	LF	LD
pięć	distractor	HF	LD
liść	distractor	LF	LD
szyb	distractor	LF	LD
twą	distractor	LF	LD
mąż	distractor	HF	LD
grom	distractor	LF	LD
tyś	distractor	LF	LD
klej	distractor	LF	LD
miś	distractor	LF	LD
małp	distractor	LF	LD
tył	distractor	LF	LD
sęk	distractor	LF	LD
sąd	distractor	LF	HD
cel	distractor	HF	HD

Word	Class	WF	PND
noc	distractor	HF	HD
ląd	distractor	LF	HD
rząd	distractor	HF	HD
czym	distractor	HF	HD
list	distractor	HF	HD
kim	distractor	HF	HD
dać	distractor	HF	HD
tąd	distractor	LF	HD
kąt	distractor	LF	HD
dom	distractor	HF	HD
rąk	distractor	LF	HD
hej	distractor	HF	HD
bóg	distractor	HF	HD
dasz	distractor	HF	HD
ruch	distractor	HF	HD
ból	distractor	HF	HD
masz	distractor	HF	HD
sen	distractor	HF	HD
rok	distractor	HF	HD