Vowel length contrasts in deep learning: Generative Adversarial Phonology and duration

Wing-Sze Kat

The Chinese University of Hong Kong

1 Introduction

Recent advances in artificial intelligence have significantly enhanced phonological studies, allowing for a more precise modeling of the human encoding process that organizes sounds and connects them to cognitive functions. An important aspect of this encoding is the contrast in vowel length, where the duration of a vowel can change a word's meaning. Languages exhibiting vowel length contrast may use either vowel quantity (duration) or vowel quality (formant frequencies) to signal distinctions. For instance, in Cantonese, the contrast between /äː/ and /ɐ/ is common, correlating with systematic changes in vowel quality and/or duration in East and Southeast Asian languages. This paper uses Cantonese, Vietnamese, and Thai to examine vowel length contrasts within the context of tone.

In prosodic phonology, the ambiguity of vowel length in prosodic expressions—especially within bimoraic structures (Vː or VV) in the Optimality Theoretic framework—arises from various factors, including durational variation, shifts in vowel quality, and language-specific rules. Furthermore, vowel length is intricately related to syllabic constituents and rime within non-linear phonology, which are CV and CVC. Vowel length remains an understudied topic, particularly concerning its definition and its influence on machine-generated Voice Onset Time (VOT) and Voice Offset Time (VOFT), which reflects the relationship among syllabic constituents.

To model these interactions, we utilize generative adversarial network (GAN) frameworks adapted for phonetic synthesis—to simulate the formation and phonological constraints of these systems by manipulating two categories: vowel quality and duration. CiwGAN facilitates the analysis of factors such as intensity and other elements like F1 (which represents vowel height), among others. By generating synthetic speech data, ciwGAN reveals how vowel length modulates tone realization across different qualities and creates long vowels from short vowels to imitate the language acquisition process.

2 The core research question in phonology and 'acquisition' way in categorical InfoWaveGAN

Previous research in deep learning applied to phonology (Beguš, 2021) shows that methods like GANs and CNNs effectively interpret identity-based patterns and model generative phonological processes, such as reduplication, in language acquisition.

Within the categorical InfoWaveGAN (ciwGAN) framework, the research question in this paper is how this ciwGAN can model and analyze phonological patterns in language acquisition, specifically to simulate how infants learn to derive long vowels from short ones. It focuses on controlling and conditioning the generation process based on discrete linguistic features. The ciwGAN architecture integrates the Deep Convolutional GAN framework designed for audio data with categorical variables, it

* Acknowledgement: This research was not funded by any grants. However, I would like to express my heartfelt gratitude to my family, my informants, and the audience of AMP 2025. I extend my thanks to Mr. Ma Kai-lun Donovan, Professor Gašper Beguš, Professor Cheung Yam-leung Lawrence, Professor Lau Chaak-ming, Professor Lee Tsz-Ming Tommy, Professor Mok Pik-Ki Peggy, Dr. Wong Tak-Sum Sam, Dr. Lai Yik-Po, Dr Chao Lip-Yan Felix, and Mr. Au Yeung Sing-Leung Edmond, the school principal of Cognitio College (Hong Kong), for their valuable comments and / or any assistance. Any errors are my own.

employs a binomial distribution for its latent code and uses sigmoid cross-entropy for training. The network is designed to learn identity-based patterns, like vowel length contrasts.

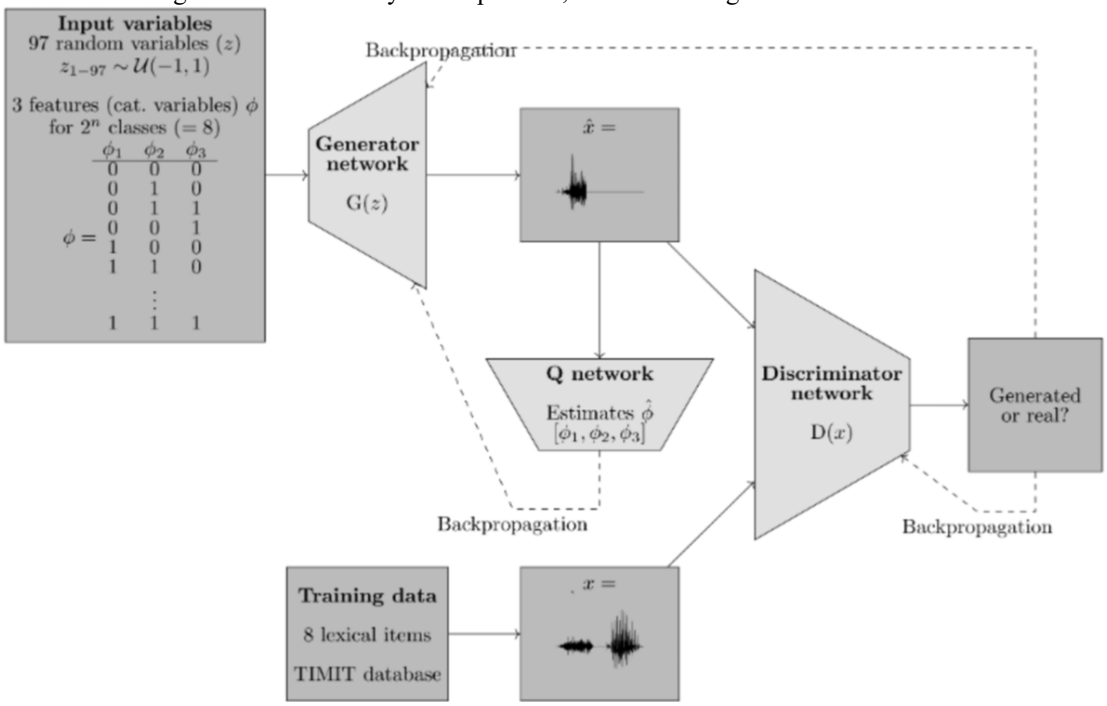


Figure 1. The mechanism of ciwGAN as proposed in Donahue et al. (2019)

This study investigates how meaningful representations emerge in Convolutional Neural Networks (CNNs) and the correlation between latent space variables and vowel length contrasts. Manipulation involves encoding these variables, conditioning the generator and/or discriminator on these features, and allowing the network to generate outputs reflecting the desired categorical properties. The network's ability to generate long vowels from short ones demonstrates its understanding of this categorical distinction, akin to human learners. The generator creates audio from random noise, while the discriminator distinguishes real from generated audio, maximizes the mutual information between parts of the generator's input and the generated data, enabling the model to learn meaningful and controllable features.

3 Materials

3.1 *The ciwGAN Model (Donovan Ma, 2025)* This study investigates the long-short vowel contrast as a categorical distinction, focusing on both its duration and vowel quality within the boundaries of lexical meaning. Given this background, there is a pressing need for research employing deep learning approaches. Indeed, deep learning models can analyze and model vowel contrasts, providing valuable insights into how these categorical distinctions are learned and represented in both human cognition and artificial systems. In our ciwGAN framework, we utilized binary duration labels (0 for short and 1 for long) based on the assumption that duration functions as an independent categorical variable. Furthermore, the generator within this model is designed to learn to control timing while maintaining consistent audio quality. This approach underscores the necessity for the GAN to differentiate not only by timing but also by the inherent phonetic characteristics associated with each vowel type.

3.2 *Training Data* The code testing environment consists of Python 3.12.3, TensorFlow 2.17.1, and a GPU (Nvidia 3060 laptop with 6 GB VRAM). Initially, we migrated the legacy TF1 codebase to TF2's tensorflow.keras framework and implemented WGAN GP loss, while also optimizing the upsampling process. Moreover, recordings of the training data (N=629) were made using Audacity. The audio was originally sampled at 44.1 kHz and then downsampled to 16 kHz. Specifically, the training data includes

Cantonese (Hong Kong, N=62), Vietnamese (all dialects, N=353), and Thai (Suphan Buri, N=533), all of which are tonal languages. The trend of the system in bellow tells every batch's operation.

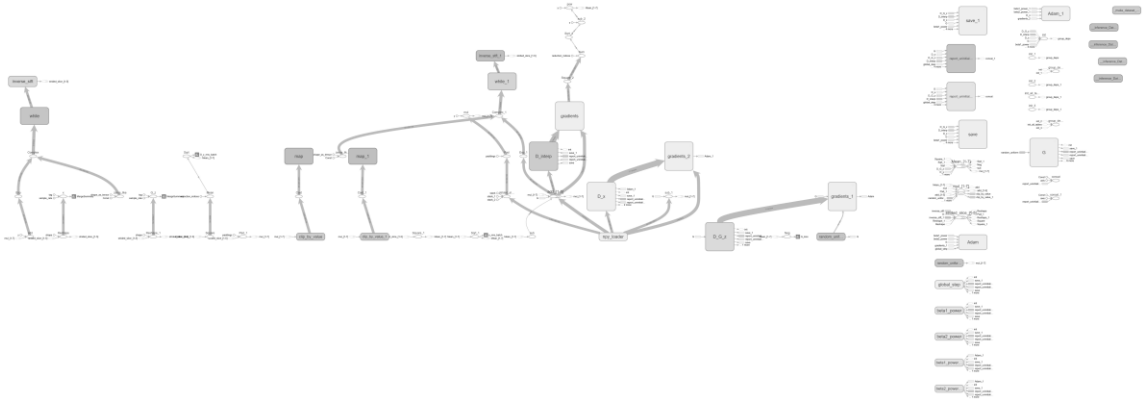


Figure 2. Batch normalization

The study explores how meaningful representations of identity-based patterns emerge in CNNs, specifically through batch normalization. In addition, it examines the correlation between latent space variables and vowel length contrast. This exploration has broader implications for understanding and interpreting neural networks. Furthermore, the exploration of how meaningful representations of identity-based patterns emerge in CNNs, sheds light on how the latent space variables, particularly those outside of the training range, correlate with vowel length contrast in the output. Last but not least, the logistic regression model achieves 92.50% test accuracy and 96.43% cross-validation accuracy, demonstrating that vowel length of the three languages can be reliably predicted from acoustic features.

4 Materials

4.1 *Epoch and learning rate* The GAN is unsupervised, meaning that category labels are automatically added. In ciwGAN, the parameter setting is achieved by encoding vowel length as a categorical variable through feature embedding. This method involves mapping categorical variables (i.e., vowel quality and duration) to dense vectors that are utilized in the generative process. In data representation, each vowel token is annotated with a length feature. To generate a long vowel from a short vowel, one simply changes the input categorical variable from short to long while keeping other features constant. Thus, latent space traversal modifies the latent representation to reflect changes in categorical variables, enabling controlled generation.

Furthermore, in the context of ciwGAN, an epoch on TensorBoard refers to a complete pass through the entire training dataset. During each epoch in our case, the model processes all available training data to update its parameters based on the calculated loss. This iterative process allows the model to gradually improve the quality and realism of the generated images. Training typically requires multiple epochs, as this enables the model to refine its performance over time, with the duration and effectiveness of each epoch dependent on dataset size and model complexity.

Finally, the results after 100 epochs indicate promising insights into the model's performance, achieving intensity similarity, which stands at 56% after normalization, suggesting partial retention of amplitude dynamics, likely due to limitations of the Griffin-Lim algorithm, which affects loudness detail. The overall weighted similarity score of 40% intensity weighting, signifies the model's readiness as a proof-of-concept. This notably highlights the successful modeling of temporal structures while identifying areas for improvement in amplitude dynamics, thereby underscoring the feasibility of phonetic GANs for applications in speech synthesis.

4.2 *Acoustic analysis using spectrogram*

4.2.1 *Spectral amplitude: intensity as the primary cue* The summary of acoustic features reveals two primary cues in vowel discrimination, with intensity as the foremost cue. Stevens (1998) noted that "the human ear is sensitive to fluctuations in sound pressure, and most microphones respond to

pressure variations."

In Thai phonetics, the relationship between formant frequencies and intensity is crucial for distinguishing vowel length. Long vowels are characterized by significantly lower first formant (F1) values, indicating a more open articulation, which is further enhanced by their greater intensity. This means that long vowels not only sound more open but are also much louder, creating a striking contrast with short vowels, which have higher F1 values and lower intensity. In our study, we observed a significant difference of 15.84 dB between short and long vowels in Thai, which supports the notion that vowel length is distinguished primarily by duration. This acoustic distinction is accompanied by a substantial effect size of d=4.43, highlighting the pronounced difference. Statistically, the finding is highly significant ($p < 0.0001$***), indicating that short vowels are approximately 16 dB quieter than their long counterparts. Specifically, long vowels measured at -67.9 dB, while short vowels were recorded at -83.7 dB. This contrast emphasizes the acoustic significance of vowel length in Thai phonetics, illustrating how both formant frequencies and intensity work together to create perceptual distinctions. Furthermore, the findings on Fourier Transform Analysis highlight a strong negative correlation between mean intensity and F1 (-0.561) in long vowels, suggesting that increased vocal effort leads to greater vocal tract constriction. Long vowels also exhibit stronger intensity-formant coupling compared to short vowels. Negative correlations between intensity and the F1/F2 ratio indicate spectral flattening at higher intensities, while positive correlations (~0.15-0.16) point to vowel space expansion with increased vocal effort. This underscores the complex interactions among multiple acoustic dimensions that contribute to contrasts in vowel length.

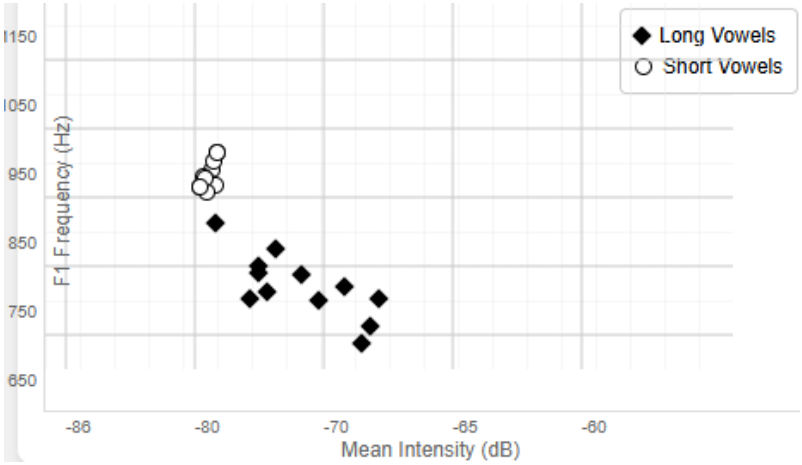


Figure 3. Mean Intensity vs F1 Frequency Scatter Plot in Thai

Understanding how intensity changes affect formants is essential in speech signal analysis, as these fluctuations can enhance the clarity and prominence of formants, thereby improving speech intelligibility. Variations in intensity can alter the spectrum shape, shifting formant frequencies and influencing the perceived quality of speech sounds. Additionally, higher intensity levels may introduce non-linear effects that affect how listeners discern the intended message. Overall, recognizing the impact of intensity on formants is crucial for effective communication and for the development of better speech synthesis models. An analysis of intensity after normalization for the GAN model, trained over 100 epochs on long vowels, further illustrates this difference. The real mean intensity is -37.70 dB, while the raw generated mean intensity is significantly quieter at -79.37 dB—42 dB too quiet. After normalization, the generated mean intensity improves to -53.08 dB but remains 15 dB quieter than the real mean, with a similarity rate of just 59%. This indicates a noteworthy divergence in intensity characteristics between real and generated outputs, highlighting the model's need for enhancements in replicating the targeted phonetic input's intensity levels. The raw generated audio reveals considerable audibility issues, being roughly 42 dB quieter, equating to about 100 times lower in amplitude than real audio. This discrepancy largely stems from the Griffin-Lim phase reconstruction process, which often results in the loss of critical amplitude information.

In Vietnamese, the #VT structure similarly relies on intensity to differentiate long and short vowels. While Vietnamese utilizes both formant and intensity cues for vowel distinctions, its acoustic patterns

exhibit notable differences compared to Thai. The expected relationship between F1 and intensity in Vietnamese still requires thorough testing; preliminary findings suggest that short vowels may demonstrate longer voiceless offsets instead of pronounced intensity differences. The model shows proficiency in generating Vietnamese vowels, accurately producing qualities such as /ɛ/ for long vowels with the correct timing. In contrast, for short vowels, the generated qualities like /e/ reveal slight timing inaccuracies, measuring 15 ms instead of the ideal 7.5 ms. This indicates a strong capability in vowel generation, needing only minor adjustments for short vowel timing. These results underscore the model's effectiveness in generating Vietnamese vowels. The findings reveal an asymmetric performance: long vowels achieve 100% accuracy, while short vowels reach only 50%. This suggests that the model differentiates between long and vowel quality, rather than focusing solely on duration.

Moreover, a logistic regression analysis in Cantonese reveals an intensity coefficient (coef = +2.62) identifying intensity as a primary acoustic cue for distinguishing vowel length. However, Cantonese demonstrates a different pattern; it relies more heavily on tonal variations and syllable structure, often leading to weaker relationships between intensity and formants in distinguishing vowel lengths. In this language, vowel length contrasts are frequently conditioned by syllable type rather than by acoustic cues such as intensity or formant differences. Thus, while Thai exhibits a robust integration of formant and intensity in delineating vowel lengths, Vietnamese and Cantonese reveal language-specific strategies that reflect their unique phonological systems. This diversity underscores the multifaceted nature of vowel perception across these tonal languages. Additionally, the first tone in Cantonese, labeled /55/, is notably loud, measured at approximately -26.04 dB, making it the loudest among all tones. Typically, high tones are 8-9 dB louder than low tones due to greater vocal fold tension. While this finding does not directly relate to the distinctions between long and short vowels, it is noteworthy that many words containing the vowel pairs /äː/ and /ɐ/ are associated with the first tone /55/. Understanding these intensity dynamics enriches our comprehension of language development and the cognitive processes involved in syllable organization. Lei (2007) highlights the limited literature on infants' sensitivity to the internal organization of syllables, particularly regarding consonant and vowel contrasts. This gap in understanding emphasizes the importance of further exploration into how humans perceive and process syllabic structures.

In conclusion, the intensity of vowels serves as a significant acoustic cue for distinguishing vowel length in Thai, Vietnamese, and Cantonese, although each language demonstrates unique patterns. In Thai, long vowels are considerably louder than short vowels, with an intensity difference of approximately 15.84 dB, emphasizing how Thai utilizes both intensity and lower F1 values to convey vowel length distinctions effectively. Vietnamese also relies on intensity, but its relationship with formant frequencies remains to be fully explored, as initial findings suggest that short vowels may exhibit longer voiceless offsets rather than pronounced loudness differences. In contrast, Cantonese prioritizes tonal variations and syllable structure, resulting in weaker correlations between intensity and vowel length; although intensity is recognized as an important factor, the language's vowel contrasts are predominantly influenced by syllable type. This comparison highlights the diverse strategies that these tonal languages employ to achieve phonological distinctions, revealing the multifaceted nature of vowel perception.

4.2.2 *Spectral shape: formant 1 as a secondary cue for vowel quality* Another cue is vowel quality, specifically the F1. There is no universal phonological principle that makes vowel length and vowel quality mutually exclusive as contrastive features. Many languages utilize both features, either independently or in interaction. Here, the overall separation is measured at 218.9 Hz, with an effect size of $d = 2.19$, also classified as a huge difference. The statistical significance is again very high ($p < 0.0001$***), demonstrating that short vowels have significantly higher F1 values, indicating they are more open. For reference, long vowels are recorded at 887 Hz, whereas short vowels are at 1106 Hz.

In examining vowel charts, we observe distinct scenarios across these languages. Thai typically features pairs of vowels for each quality, which maintain similar positions in the F1-F2 space while differentiating in duration. Vowel length is significant in Thai, serving as a major contrastive feature that allows for both short and long vowels across most qualities. The vowel pairs in Thai are usually represented for each quality (e.g., /a/ vs. /aː/), often occupying the same position in the F1-F2 space but differing notably in their duration (i.e., vowel length).

Vietnamese, in contrast to some other languages, primarily differentiates vowel quality, with certain vowels being intrinsically longer or shorter, although this length does not serve as a phonemic distinction.

In Vietnamese phonetics, long vowels feature a higher F1, approximately +108 Hz compared to short vowels. From a phonetic perspective, some vowel qualities are more easily discerned by duration; for example, high vowels are generally shorter than low vowels across many languages. Nevertheless, the F1 distinction, which reflects vowel openness (i.e., vowel height), remains a crucial characteristic in Southeast Asian languages, particularly in Vietnamese. The analysis of Vietnamese charts shows quality contrasts, indicating that while some vowels have intrinsic lengths, this does not imply a phonemic contrast. Overall, the phonetic strategies observed in Vietnamese underscore diverse approaches to vowel discrimination, emphasizing the importance of both spectral and length features within its distinct phonological system.

In Cantonese, vowel quality serves as the primary means of contrast, while vowel length is not a phonemic feature; thus, its vowel chart reflects only quality contrasts. In Luo(2019)'s paper, it was noted that Lee (1983, 1985) found, and Shi and Liu (2005) later confirmed, that even at comparable speech rates, the duration ranges of long and short vowels in a vowel pair do not overlap, despite significant spectral overlap (Zee, 2003). This substantial spectral overlap indicates that vowel quality alone is not the distinguishing factor for contrast.

When we compare these three languages, vowel openness emerges as the most substantial cue. Short vowels demonstrate significantly higher F1 values, suggesting they are more open (higher F1 with /e/, /ɛ/, /a/, /ɔ/ in Thai, etc.) and front, with short vowels measured at 1106 Hz and long vowels at 887 Hz, indicating a notable difference. The effect size of this distinction is d = 2.19, suggesting that, like Thai, long and short vowels in Vietnamese also differ in quality.

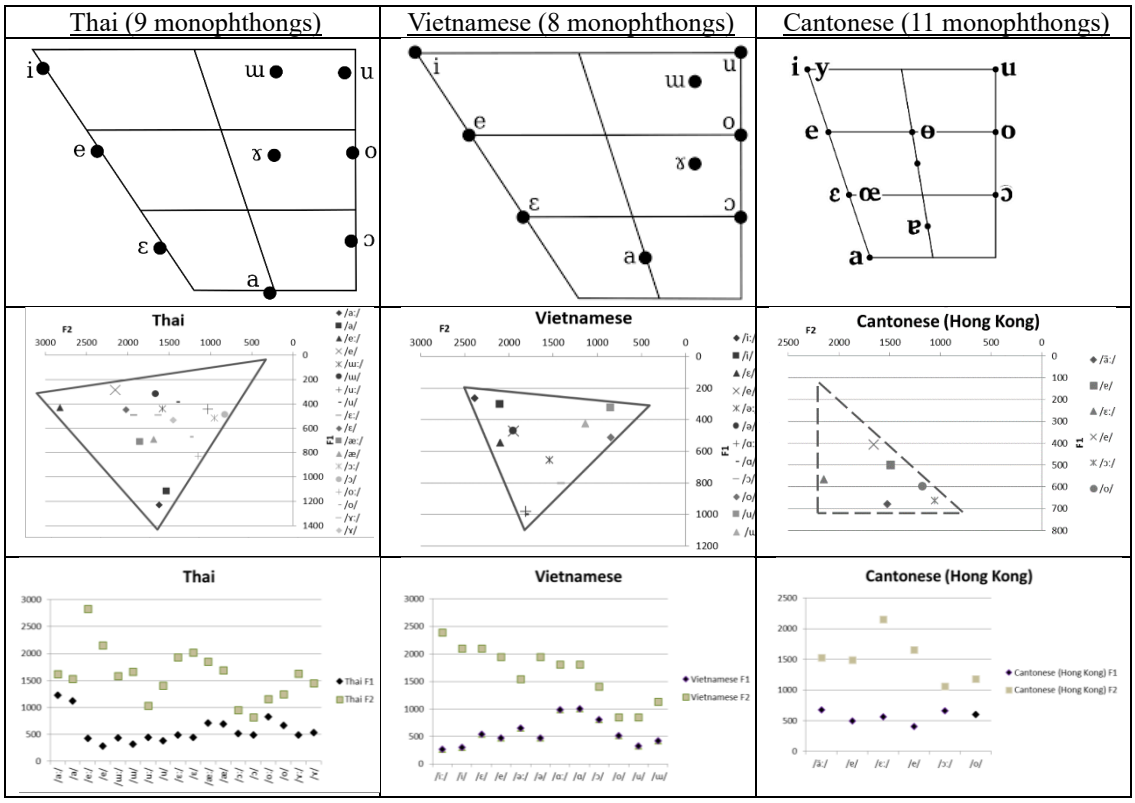


Figure 4. Vowel charts of monophthongs and Formants of contrasting vowels

The term “vowel openness” can be somewhat ambiguous in phonological discussions, as vowel charts may vary significantly across languages. To address this variability, representing vowel systems through formant extraction offers a more objective and quantifiable depiction of vowel quality. In our phonetic experiment using Praat, as illustrated in Figure 4, we observed that the distribution of vowel height and backness differs across the three languages under investigation. Specifically, the contrasting vowels in Cantonese are realized with relatively low formant values, indicating lower tongue positions. In Vietnamese, back vowels are also produced with low tongue positions, the [-mid] long vowels (/iː/, /ɯ/, /ɑː/) in the #TV and #VVT syllable structures display higher F1 values to align with their corresponding

short vowels, whereas in Thai, back vowels are articulated with higher tongue positions, the [+mid] long vowels (/ɛː/, /ɤː/, /ɔː/) generally exhibit higher F1 values to correspond with their short vowel counterparts. These findings highlight the importance of using acoustic measures, such as formant frequencies, to accurately compare vowel systems across languages and to address why long vowels are closer to short vowels. In summary, the ciwGAN aligns well with the vowel charts, effectively distinguishing both vowel quality and duration.

5 Discussion

5.1 *Parallels between L1 language acquisition and language pathology in ASD: insights from spectrogram analysis* In language acquisition, vowel acquisition occurs at an early stage in the L1 development of both Cantonese and Vietnamese (Phạm et al., 2019). In the GAN-generated data, [front, long] vowels (e.g., /i/ of Vietnamese and /ɛ/ or /ɛː/ of all three languages) are first observed at step 50,024. This observation aligns with the vowel chart's classification of long vowels, which depends on the distance between the F1 and F2. The data on vowel openness reveal that long vowels tend to have higher F1 values and longer durations. Furthermore, the frequency gap in F2 between /ɛ/ and /æ/ is significantly greater than that between /ɪ/ and /ɛ/, highlighting a distinct acoustic separation among these front vowels. Additionally, the Cantonese /ɛ/-/e/ pair of front vowels is also acquired in L3 (Luo et al., 2019).

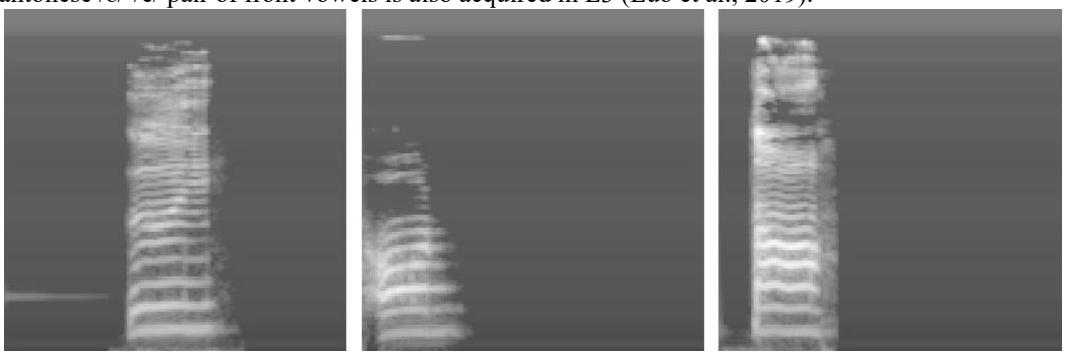
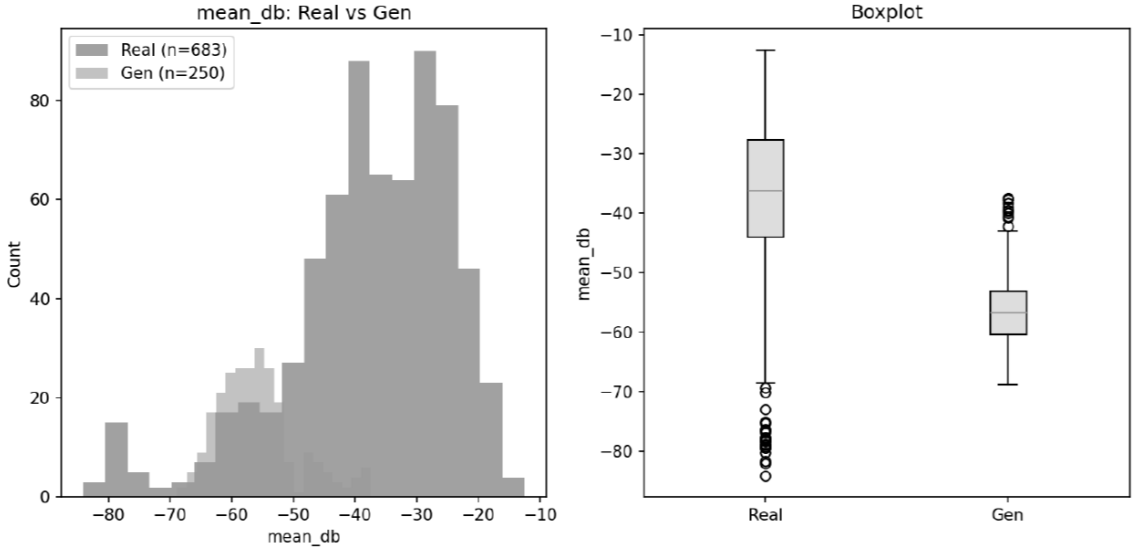


Figure 6. Generated sounds at step 50,024.

However, the imitation of learners' behavior by ciwGAN is not perfect, as generator network violates the intensity distribution. The generated vowels may not fully replicate the loudness and quality characteristics found in the natural recordings.



(a) (b)

Figure 7. (a) histogram: the intensity distributions of loudness between two groups: real audio and generated short vowels, and (b) boxplot: a detailed statistical comparison of loudness between real and generated audio samples.

The plot compares the intensity distributions of loudness between two groups: real audio and generated short vowels. The charcoal gray bars represent the intensity distribution of 683 real audio files, showcasing the natural loudness patterns present in authentic speech. The French grey bars illustrate the intensity distribution of 250 generated short vowel files after normalization, indicating how the model's output aligns with or deviates from real audio. This comparison highlights differences in loudness profiles, providing insights into the quality and accuracy of the generated sounds in relation to authentic phonetic characteristics.

The left panel features an overlay histogram that compares intensity distributions between real and generated audio samples. The X-axis represents intensity in decibels (dB), while the Y-axis indicates frequency, showing how many samples correspond to each intensity level. This visual representation highlights the degree of overlap between the two distributions, providing insights into how closely the generated vowels mimic the natural loudness patterns found in authentic speech. Analyzing this overlap can reveal areas where the model's imitation succeeds or falls short.

For peak location, this indicates where most samples cluster in terms of intensity. The real audio peaks around -36 dB, while the generated audio peaks at about -56 dB. This suggests that the generated audio is, on average, ~20 dB quieter than the real samples. For the Spread (Width), this refers to the variability in loudness. The real audio has a wider spread, ranging from approximately -60 to -20 dB, indicating a dynamic range with significant volume fluctuation. In contrast, the generated audio has a narrower spread, ranging from -70 to -40 dB, showing less variation in loudness. This implies that the generated audio lacks the natural volume fluctuations characteristic of authentic speech. For the Overlap, there is a small amount of overlap between the two distributions, indicating that the generated and real samples can be distinguished based on loudness alone. This suggests that while the model produces recognizable sounds, there are notable differences in their loudness profiles.

For median comparison, the median loudness for real audio is -36.24 dB, while the generated audio has a median of -56.78 dB, thereby creating a gap of 20.54 dB. This indicates that the generated audio is consistently quieter than the real samples. Furthermore, regarding Box Height (IQR), the taller box for the real audio represents greater variability in loudness, whereas the shorter box for the generated audio reflects reduced variability. This suggests that real speech encompasses a richer dynamic range.

In terms of whisker range, the whiskers for the real audio extend from approximately -65 to -15 dB, indicating a range of 50 dB. In contrast, the generated audio ranges from about -75 to -40 dB, showing a narrower range of 35 dB. This implies that the generated audio lacks the extreme quiet and loud moment’s characteristic of natural speech.

We find that intensity serves as a significant cue in the speech of speakers with autism spectrum disorder (ASD), individual variation in ASD characteristics leads to selective structural language impairments. Although there is no research specifically on ASD in Thai children, a study by Chanchaochai (2019) explores the role of intensity in experiment preparation. Recent studies (Schaeffer et al., 2023) investigating the impact of intensity and duration on speech flow have yielded mixed results. For instance, Hubbard & Trauner (2007) emphasized significant variations in vowel intensity between participants with and without autism, underscoring the relevance of intensity in speech production. Similarly, Olivati et al. (2017) observed comparable disparities in utterance intensity.

In addition to intensity, vowel quality is another important cue for language acquisition. Hu et al. (2025) demonstrated that a compact and diverse set of non-Mel-Frequency Cepstral Coefficients (MFCCs) and selected spectral features, such as F1 in CV structures and duration, effectively characterize speech abnormalities in verbally fluent individuals with ASD. This highlights the potential of data-driven models to complement clinical assessments and enhance our understanding of speech-related manifestations.

5.2 *Consonants and Vowels in VOT imitation* Educator Hallie Kay Yopp (1992) noted a relationship between consonants and vowels; however, her observation lacked experimental support. She recalled that “Henderson (1982) noted how phonemes are influenced by their phonological context.” Beguš (2022) refers to this as “minor local dependencies.” This phenomenon arises because the articulatory positioning and tension required to produce high vowels (/i/, /u/) can affect the timing of the vocal folds' onset of vibration, leading to a longer Voice Onset Time (VOT) compared to lower vowels.

VOT may be influenced by vowel height, with high vowels potentially causing a delay in voicing.

Through ciwGAN, we can analyze how syllabic structures are generated and investigate how vowels influence consonants. On syllable structure effects, the position of the vowel within the syllable can affect the salience of duration and quality cues. For example, vowels in open syllables may have more variable duration, while those in closed syllables may be more stable. This paper will test both CV (consonant-vowel) structures and VC (vowel-consonant) structures. # means boundary, T represents unaspirated obstructs /p t k/ and D represents voiced obstructs /b d g/.

Table 1. CV and VC structures

	Thai	Vietnamese	Cantonese
#TV	✓	✓	
#T^hV	✓	(✓)[1]	
#DV	✓	✓	
#VT	✓	✓	✓
#VVT		✓	
#VD	✓		

Infants and young children are sensitive to both spectral (quality) and temporal (duration) cues from an early age. In Thai, long vowels have a mean VOT of approximately 26.4 ms, while short vowels have a mean VOT of about 19.1 ms, resulting in a difference of 7.3 ms ($p < 0.05$, Cohen's $d = 0.242$). This small but significant VOT difference in Thai emphasizes a clear distinction between long and short vowels.

In Vietnamese, the VOT for long vowels is recorded at 0.00 ms, while the VOT for short vowels is approximately 11.25 ms, so it plays a critical role in distinguishing vowel length contrasts in Vietnamese: short vowels exhibit significantly longer VOT than long vowels, with a mean difference of approximately 26.56 milliseconds; long vowels average 21.34 ms, while short vowels reach 47.90 ms. This inversion illustrates how Vietnamese speakers compensate for reduced vowel duration by increasing aspiration and breathiness. As a result, VOT emerges as a key acoustic cue in the Vietnamese phonetic inventory, emphasizing the complex interplay between duration and aspiration in achieving clear vowel contrasts.

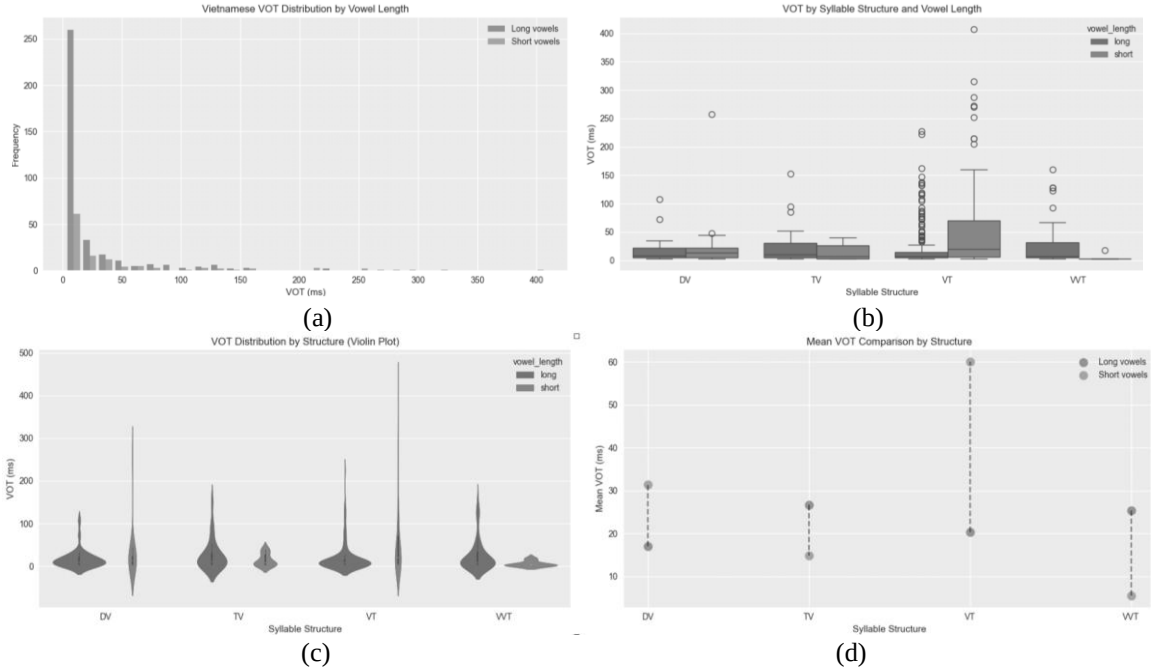


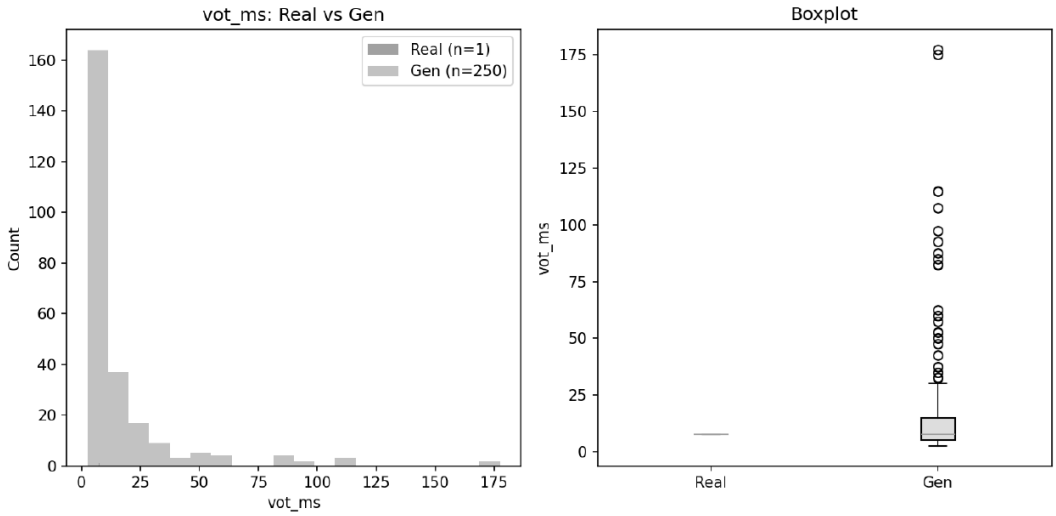
Figure 8. The relationship between VOT and the opposition of long and short vowels in Vietnamese: (a) VOT distribution by vowel length, (b) VOT by syllable structure and vowel length, (c) VOT distribution by structure (Violin Plot) and (d) Mean VOT comparison by structure.

[1] In Vietnamese, the #T^hV structure involves an aspirated voiceless stop, specifically represented by /t^h/. This contrasts with the unaspirated voiced /d/ and the voiceless unaspirated /t/. However, we did not include it in our experiment.

This research that employs ciwGAN fills the research gap of how native speakers of Cantonese, Thai and Vietnamese perceive syllabic constituents under different phonetic conditions, concerning language acquisition for vowels. A crucial question regarding GAN is whether the audio tracks produced by ciwGAN closely resemble the original syllables. Additionally, we must evaluate the accuracy of the generated output, focusing on VOT for stops, which is defined as the interval between the release of the stop and the onset of vocal cord vibration in the following vowel. Our approach models speech acquisition as a dependency between a latent space (X) and the generated speech data within the GAN framework. This structure lays the groundwork for categorizing contrasts. Testing vowel quality will help elucidate how this local factor affects VOT, serving as an effective method for evaluating the imitation accuracy.

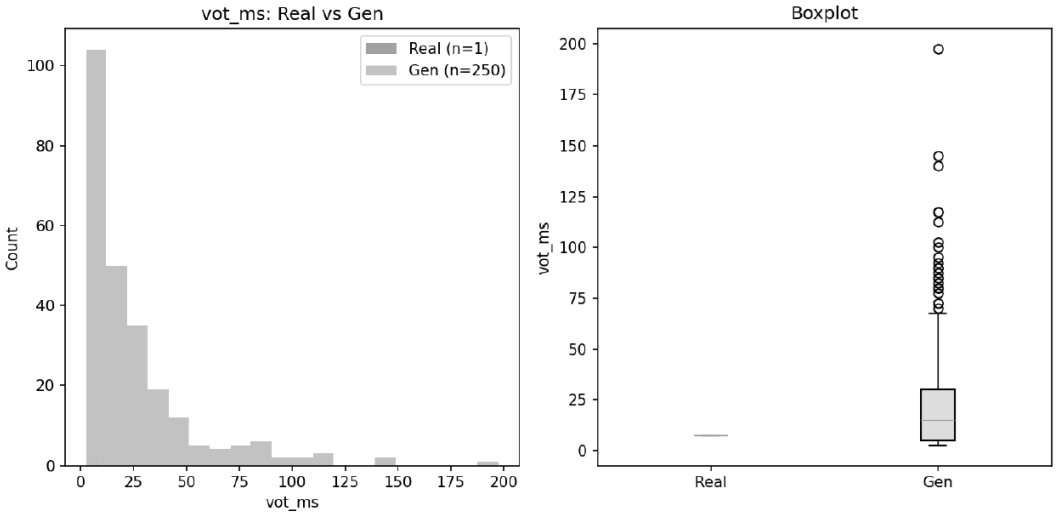
Further suggested phonetic cues in this research include examining the periodic vibration of adjacent vowels and the maximum intensity of the syllabic structure combination nor local dependencies, such as the periodic vibration of adjacent vowels, can influence the duration of VOT. This influence can be measured by analyzing the proportion of vowels and consonants within various syllabic constituents.

For timing measurements, we will employ a (a) histogram and (b) boxplot structure, with the X-axis representing VOT in milliseconds (ms). Small values (0-10 ms) indicate tight consonant-vowel timing, while larger values (100 ms and above) reflect a longer delay between the consonant and the vowel.



(a) (b)

Figure 9. Long vowels plot



(a) (b)

Figure 10. Short vowels plot

These results indicate that the model has successfully learned the temporal structures associated with long vowels, effectively capturing the nuances of phonetic timing. However, the systematic errors observed in the timing of short vowels suggest that the model may be misinterpreting the relationship between duration and vowel quality. In a current study (Nguyen-Phuoc et al., 2025), a fully separated VOT range for the three alveolar stops was observed in 4-year-old children, while some children in other age groups exhibited overlapping VOT values between the stop categories. Although the paper discusses the relationship between VOT and vowel onset F0, it overlooks the timing of short vowels with differing phonetic patterns. Our model faces significant challenges with short vowels, likely because it does not accurately capture the relationship between vowel length and its corresponding timing dynamics. Specifically, the model struggles to differentiate the rapid timing typically associated with short vowels from that of longer or different vowel types. This issue may also stem from the varying vowel qualities in pairs such as /ɛ/ - /e/, /ɔ/ - /o/, and /u/ - /ɯ/, complicating the model's ability to accurately represent short vowel timing. Additionally, another study (Trần et al., 2019) confirmed that Vietnamese vowel duration contains information about the place of articulation of the following coda stop in #VT and #VVT structures of Vietnamese.

In conclusion, we evaluate how the network learns to categorize vowel length through VOT, focusing on its ability to imitate human speech and support language acquisition. The results indicate a mixed performance in VOT imitation. For long vowels, the model successfully demonstrates accurate imitation, as evidenced by the exact match in timing at 7.5 ms for both real and generated data, reflecting a solid understanding of the temporal structure associated with long vowels. In contrast, the imitation for short vowels presents significant challenges; the generated vowel peaks at 15 ms instead of the expected 7.5 ms, leading to a notable timing error. This suggests that, while the model has made commendable progress, it conflates the timing of short vowels with different phonetic patterns. Overall, the model captures the timing dynamics effectively for long vowels, but further improvements are needed to enhance its accuracy in modeling short vowel timing.

6 Conclusion

In conclusion, our computational linguistics experiment marks a significant advancement in understanding vowel length contrasts, particularly in the transformation of short vowels into long vowels. While vowel duration is identified as a crucial phonemic distinction in Thai, in Cantonese and Vietnamese, vowel quality takes precedence. Interestingly, intensity emerges as the primary cue for distinguishing vowel length across these languages. This duality not only underscores the complexity of phonological alternation but also bridges a long-standing research gap in the debate between symbolism—represented by V: or VV—and connectionism within cognitive science. This is especially pertinent in the context of autism spectrum disorder, where understanding linguistic processing can have significant implications.

Moreover, our results contribute valuable insights into the cognitive mechanisms underlying language acquisition and phonological processing. By revealing how systematic manipulation of vowel characteristics can lead to the generation of long vowels, we underscore the importance of integrating computational models with linguistic theory. Ultimately, this research paves the way for further exploration into the dynamics of vowel contrasts, enriching our understanding of linguistic structures and their cognitive representations across diverse languages.

7 Reference

Beguš, Gašper. (2020). Generative Adversarial Phonology: Modeling Unsupervised Phonetic and Phonological Learning with Neural Networks. *Frontiers in Artifical Intelligence*. 3:44.

Beguš, Gašper. (2021). Identify-Based Pattern in Deep Convolutional Networks: Generative Adversarial Phonology and Reduplication. Cambridge, MA: *MIT Press*, p. 1180–1196.

Beguš, Gašper. (2022). Local and non-local dependency learning and emergence of rule-like representations in speech data by deep convolutional generative adversarial networks. Computer Speech & Language 71 (2022).

Boersma, Paul & Weenink, David. (202024). *Praat: doing phonetics by computer* (Version Praat (version 6.4.27) [Computer program]. Retrieved March 15, 2024, from http://www.praat.org/

Chanchaochai, Nattanun. (2019). Language Profiles of Thai Children With Autism: Lexical, Grammatical, And Pragmatic Factors. Publicly Accessible Penn Dissertations. 3562.

Donahue, Chris. (2019). fiwGAN (Featural InfoWaveGAN): Lexical Learning in Generative Adversarial Phonology. Retrieved July 17, 2025, from https://github.com/gbegus/fiwGAN-ciwGAN

Chuanbo Hu, Jacob Thrasher, Wenqi Li, Mindi Ruan, Xiangxu Yu, Lynn K. Paul, Shuo Wang, and Xin Li. (2025) Speech pattern disorders in verbally fluent individuals with autism spectrum disorder: a machine learning analysis. *Front. Neuroinform.* 19:1647194. doi: 10.3389/fninf.2025.1647194

Henderson, Lillian. (1982). Orthography and word recognition in reading. *Academic Press*.

Hubbard K & Trauner DA (2007). Intonation and emotion in autistic spectrum disorders. *Journal of Psycholinguistic Research* 36(2):159-173.

Kao, Diana. (1971). The structure of the syllable in Cantonese. The Hague, The Netherlands: Mouton.

Lee, Thomas. Hun.-tak. (1983). The vowel system in two varieties of Cantonese. *UCLA Working Papers in Phonetics*,57, 97–114.

Lee, Thomas. Hun.-tak.. (1985). Guangzhou hua de yinzhi ji chang duan duili (Opposition of vowel quality and length in Cantonese). *Fangyan* (Dialect), 1, 28–38.

Lei, Margaret Ka-Yan. (2007). Discrimination of level tones in Cantonese-learning infants. Poster presented at the 16th International Congress of Phonetic Sciences (ICPhS-16), August 6-10, Saarbrücken, Germany.

Luo Jingxin & Li Guo Vivian & Mok, Pik Ki Peggy. (2019). The Perception of Cantonese Vowel Length Contrast by Mandarin Speakers. *Language and Speech*, p.1-25.

Ma, Kai-lun Donovan. (2025). ciwGAN. Retrieved January 1, 2026, from https://github.com/domna735/GAN-2025.

Nguyen-Phuoc, Tam Minh & Lee, Sue Ann. Acoustic characteristics of stop consonants in Vietnamese children and adults. *Journal of Child Language*. Published online 2025:1-19. doi:10.1017/S0305000925100123.

Olivati, Andrea, et al. (2017). Acoustic analysis of speech intonation pattern of individuals with Autism Spectrum Disorders. *CoDAS*, *29*(2), e20160081.

Phạm, Ben & McLeoda, Sharynne. (2019). Vietnamese-Speaking Children's Acquisition of Consonants, Semivowels, Vowels, and Tones in Northern Viet Nam. *Journal of Speech, Language, and Hearing Research*, Vol. 62, p.2645–2670.

Schaeffer, Jeannette ; Abd El-Raziq, Muna ; Castroviejo, Elena ; Durrleman, Stephanie ; Ferré, Sandrine ; Grama, Ileana ; Hendriks, Petra ; Kissine, Mikhail ; Manenti, Marta ; Marinis, Theodoros ; Meir, Natalia ; Novogrodsky, Rama ; Perovic, Alexandra ; Panzeri, Francesca ; Silleresi, Silvia ; Sukenik, Nufar ; Vicente, Agustín ; Zebib, Racha ; Prévost, Philippe ; Tuller, Laurice. (2023). Language in autism: domain, profiles and co-occuring conditions. *Journal of Neural Transmission* (2023)130:433-457.

Shi, Feng & Liu, Yanhui. (2005). Guangzhou hua yuanyin de zai fenxi (Revisit vowels in Guangzhou Cantonese). *Fangyan* (Dialects), 1, 1–8.

Stevens, Kenneth N. (1998). *Acoustic Phonetics*. Cambridge, Massachusetts: The MIT Press.

Trần Thi Thuy Hien, Vallée Nathalie, Granjon Lionel. Effects of Word Position on the Acoustic Realization of Vietnamese Final Consonants. *Phonetica*. 2019;76(1):1-30. doi: 10.1159/000485103. Epub 2018 May 28. PMID: 29852503; PMCID: PMC6878739.

Yopp, Hallie Kay. (1992). Developing phonemic awareness in young children. *The Reading Teacher*, 45(9), 696-703.

Zee, Eric. (2003). Frequency analysis of the vowels in Cantonese from 50 male and 50 female speakers. In Maria-Josep Solé, Daniel Recasens, & Joaquin Romero (Eds.), *Proceedings of the 15th international congress of phonetic sciences*, p. 1117–1120.