Phonetic Completeness Over Prosodic Diversity: Syllable-Level Synthetic Corpus Construction for Low-Resource Penang Hokkien Speech Synthesis

Yu Liang Lai[1], Yen Min Jasmina Khaw[2], Seng Poh Lim[3], Tien Ping Tan[4]
Faculty of Information and Communication Technology,
Universiti Tunku Abdul Rahman, Perak, Malaysia[1, 2, 3]
School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia[4]

***Abstract*—This study presents the first Text-to-Speech (TTS) model for Penang Hokkien, a low-resource tonal dialect at risk of extinction. To address phonological sparsity in the collected speech corpus, we propose a two-stage fine-tuning approach that emphasizes comprehensive phonetic coverage through syllable-level synthetic augmentation while subsequently refining prosodic naturalness using real speech recordings. By supplementing a limited 45-minute real speech corpus with a 2-hour syllable-level concatenative synthetic corpus, the full dialectal inventory of approximately 2,000 unique syllable-tone combinations was encompassed. Experimental results suggest that improving syllable-tone coverage contributes substantially to intelligibility and tonal accuracy in this low-resource tonal setting. Technical optimizations, including a 600-ms cross-fading technique to mitigate boundary artifacts and numerical tone markers to reduce token sparsity, further improved model stability and synthesis quality. The final model achieved a Mean Opinion Score (MOS) of 3.92.**

***Keywords*—Low-resource language; speech synthesis; text-to-speech; data augmentation techniques**

I. Introduction

Penang Hokkien is a local variant of Hokkien (Chinese dialect) spoken in Penang, Malaysia. This dialect has evolved uniquely over time, incorporating loanwords from Malay, English, Cantonese, and Teochew, reflecting Penang's rich multicultural heritage [1]. In recent years, the usage of Penang Hokkien has declined, particularly among the younger generation, as Mandarin increasingly dominates communication within the Malaysian Chinese community. This shift places Penang Hokkien at risk of extinction if efforts are not made to preserve it for future generations [2].

One viable approach to safeguarding this dialect is through comprehensive documentation, in terms of text and speech, by building a TTS system. However, developing TTS systems for low-resource languages presents significant challenges, primarily due to the lack of high-quality paired text and speech data. As a predominantly spoken, unwritten dialect, Penang Hokkien lacks a standardized orthography and a comprehensive speech corpus. Most state-of-the-art neural TTS models require extensive high-quality datasets that are infeasible for such low-resource scenarios. Furthermore, even when limited real speech is recorded, it often suffers from phonological sparsity, failing to cover the exhaustive inventory of possible syllable-tone combinations, which results in poor performance on out-of-vocabulary (OOV) terms.

To mitigate the resource constraints, this study implements a two-stage fine-tuning approach centered on phonetic completeness leveraging the SpeechT5_TTS model architecture [3]. By establishing a dense phonetic foundation through a hybrid speech corpus and optimizing tonal representation, this work proposes a potentially scalable framework and serves as an initial case study for the digital preservation of unwritten tonal dialects.

The major contributions of this study are as follows:

1) We empirically investigate the relationship between phonetic coverage and prosodic variation in low-resource tonal TTS. Experimental results suggest that expanding syllable-tone coverage through syllable-level concatenative augmentation contributes more consistently to intelligibility and tonal accuracy than prosodic perturbation alone, particularly for out-of-vocabulary (OOV) syllables.

2) We propose and validate the use of numerical tone markers as a superior alternative to traditional diacritics for low-resource TTS. We demonstrate that this representation reduces token sparsity by approximately 35% and accelerates model convergence by 18%, providing a more stable embedding space for rare syllable units.

3) We implement a two-stage fine-tuning approach using a hybrid speech corpus with an identified optimal 2:1 synthetic-to-real data ratio. Additionally, a 600-ms cross-fading technique is introduced to successfully mitigate the boundary artifacts inherent in concatenative synthesis, balancing phonetic robustness with natural human prosody.

4) To the best of our knowledge, this study establishes the first functional TTS framework for the Penang Hokkien dialect. This provides a vital technological foundation for the digital preservation of this endangered unwritten language.

II. Literature Review

A. Low-Resource TTS Dataset Acquisition

Recent advancements in neural speech technologies and the challenges of low-resource scenarios have prompted significant changes in TTS dataset construction. Traditional approaches relying on extensive audio recordings are often impractical for low-resource languages. To address this, researchers are increasingly repurposing existing data to create effective TTS datasets.

For example, LibriTTS [4] and CML-TTS [5] were adapted from ASR datasets, using carefully designed pipelines for text and audio processing to ensure compatibility with TTS systems. These pipelines retained critical features of the original ASR datasets, such as speaker diversity, text-audio alignment, and high-quality recordings, while resolving common issues like lengthy utterances, transcription inconsistencies, and noisy audios. In contrast, projects like the CMU Wilderness Multilingual Speech Dataset [6], BibleTTS [7], and Sanskrit TTS [8] have utilised publicly available Bible readings as a foundation to construct speech corpus. Through meticulous data preprocessing, these initiatives have successfully transformed crawled data into reliable resources for speech synthesis.

Researchers have also explored the potential of other found audio sources. [9] analyzed acoustic and prosodic features of various audio recordings, identifying that broadcast news is particularly suitable for TTS tasks. Similarly, a small TTS corpus for Moroccan dialect was created using storytelling audios from YouTube, which were denoised, segmented, transcribed with ASR system, and manually normalized to ensure accuracy [10]. Furthermore, unsupervised data selection approaches have shown promise in low-resource settings. In [11], researchers used broadcast data to create a TTS dataset, achieving a high Mean Opinion Score (MOS) of 4.4 for intelligibility with only one hour of data, demonstrating the potential of combining broadcast data with selection methodologies to address resource constraints effectively.

B. Low-Resource Speech Synthesis

Various approaches were employed to address data scarcity in low-resource languages. A major direction seeks to reduce reliance on paired text-speech corpora by leveraging representation learning. These approaches exploited large pre-trained models to extract intermediate representations that capture phonetic or acoustic information without requiring explicit labels. For example, language models such as Bidirectional Encoder Representations from Transformers (BERT) [12][13] and self-supervised speech models including Vector Quantized Variational Autoencoder (VQ-VAE) [14][15] and Hidden-Unit BERT (HuBERT) [16] have been employed to learn compact acoustic-symbolic units directly from raw speech. These representations approximate phoneme sequences and serve as intermediate units for training TTS models without extensive labelled data.

Beyond representation learning, unsupervised training has also been explored to bypass the need for parallel datasets altogether. For example, [17] proposed training two separate modules, one for alignment and one for synthesis on non-parallel data, enabling TTS learning without explicit text-speech pairs. Another influential direction exploited the duality between ASR and TTS through closed-loop learning. The "speech chain" framework [18], for instance, co-trains ASR and TTS by allowing TTS-generated audio to be transcribed by ASR, with the transcriptions fed back as supervisory signals to TTS. Similarly, ASR models have been used to produce pseudo-labels for unlabelled speech [19].

Complementary to unsupervised approaches, data augmentation and cross-lingual adaptation form another practical toolkit for low-resource scenarios. Conventional augmentation such as speech perturbation, pitch shifting, and background noise addition [20][21] are widely used to artificially expand existing datasets. While these approaches play a vital role, transfer learning remains the standard approach for low-resource TTS, either used solely or in combination with other approaches mentioned before. Its effectiveness was demonstrated through pretraining on a single high-resource language [22][23]. In more recent works, its capabilities have been extended to include multilingual and multi-speaker modelling for enhanced adaptability [24]. Phonetic similarity in relation to language closeness has also emerged as a key factor in transfer learning. In [25], a hierarchical transfer learning approach revealed that phonologically close languages significantly benefit from multilingual transfer learning, highlighting the importance of nuanced approaches in addressing the challenges of TTS in low-resource scenarios.

III. Methodology

To effectively address the data scarcity and phonological sparsity inherent in the Penang Hokkien dialect, this study proposes a comprehensive methodology that bridges the gap between synthetic data engineering and progressive neural model training. The overall architecture is divided into two primary phases: (a) hybrid speech corpus construction and (b) two-stage text-to-speech (TTS) model training. The first phase utilizes a dual-pipeline approach to acquire raw data, subsequently undergoing acoustic and linguistic processing, including 600-ms linear cross-fading and numerical tone representation, to generate a phonetically-rich synthetic speech corpus and a prosodically-authentic real speech corpus. In the second phase, these corpora are leveraged through a sequential fine-tuning strategy. The process begins with a pretrained SpeechT5_TTS model ($M_{pretrained}$) [3] that undergoes intermediate fine-tuning on synthetic data to establish a phonetic foundation by learning basic tonal contours across the comprehensive syllable inventory ($M_{intermediate}$), followed by final fine-tuning on real speech to achieve prosodic refinement and human-like naturalness (M_{final}) (Fig. 1).

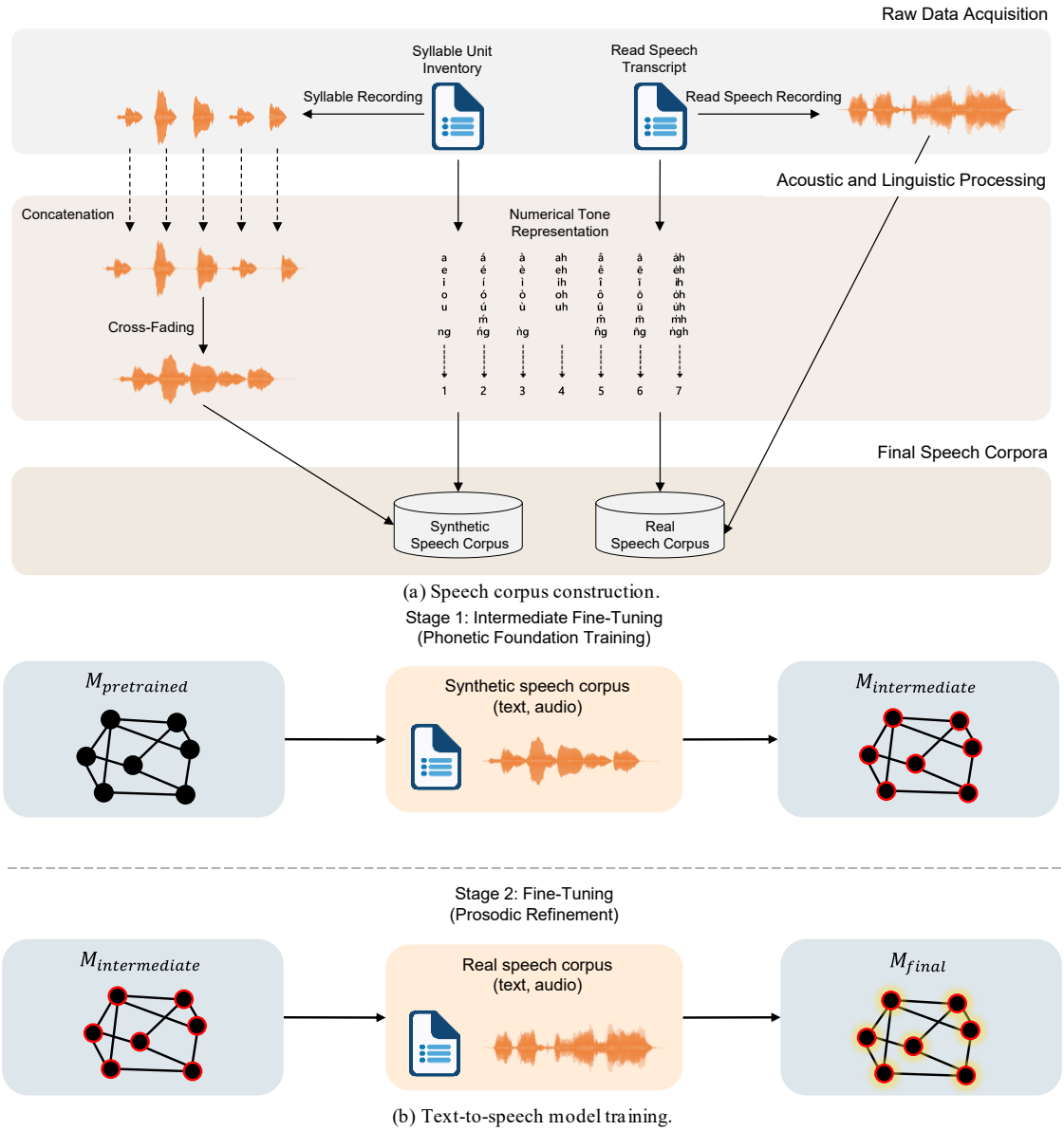

Fig. 1. The Penang Hokkien TTS framework is divided into two distinct components: (a) The speech corpus construction phase, which establishes a hybrid dataset by integrating isolated syllable units and continuous read speech via numerical tone mapping and acoustic smoothing; and (b) The text-to-speech model training phase, illustrating the progressive two-stage fine-tuning strategy used to adapt a pretrained model into a phonetically-complete and prosodically-refined final TTS system.

A. Real Speech Corpus: The Acoustic Foundation

The real speech corpus serves as the acoustic foundation for the TTS model. The transcript was sourced from documented conversations and community-driven social media platforms [26][27] to capture authentic linguistic variations. Inconsistencies in non-standard orthography were resolved using the Soundex$_{PH}$ algorithm [28] to ensure phonetic consistency. A native female speaker recorded the transcript in controlled conditions using high-fidelity equipment at 44.1 kHz to capture a broader frequency range, preserving phonetic details like the velar stop *'ak'*, which might be lost at lower sampling rates. Audios were downsampled to 16 kHz to align with the SpeechT5 architecture after undergoing silence trimming and volume normalization. After these procedures, approximately 45 minutes of high-quality Penang Hokkien real speech corpus was obtained (Table I).

TABLE I. DISTRIBUTION OF REAL SPEECH CORPUS

Statistics	Value
Number of utterances	902
Total syllable count	10507
Unique syllable count	1284
Total duration	44 min 12 sec
Minimum length of utterance	1.04 sec
Maximum length of utterance	7.22 sec
Average length of utterance	1.94 sec

B. Synthetic Speech Corpus: Bridging the Phonological Gap

To bridge the phonological gap (around 700 missing syllables in the 45-min real speech corpus), a 4-hour synthetic speech corpus was constructed (Table II). A synthetic transcript was generated through the random concatenation of romanized syllables to ensure comprehensive coverage of the approximately 2,000 unique syllable-tone combinations in the Penang Hokkien phonetic inventory. Individual syllable units were recorded in isolation by the same native speaker to maintain acoustic consistency. Syllables were utilized as the primary synthesis unit rather than traditional speech unit diphones to better reflect the tonal characteristics of Penang Hokkien while reducing computational complexity [29].

The necessity of the synthetic corpus is highlighted by the phoneme distribution analysis shown in Fig. 2 and 3. The real speech corpus exhibits a heavily skewed distribution where frequent phonemes dominate, leaving a long tail of underrepresented units. Conversely, the syllable-level concatenative approach yields a deliberately balanced distribution, ensuring that the model is exposed to the comprehensive phonetic inventory of Penang Hokkien.

TABLE II. DISTRIBUTION OF THE SYNTHETIC SPEECH CORPUS

Statistics	Value
Number of utterances	5779
Total syllable count	53481
Unique syllable count	2045
Total duration	4 hr 0 min 2 sec
Minimum length of utterance	1.65 sec
Maximum length of utterance	7.08 sec
Average length of utterance	2.24 sec

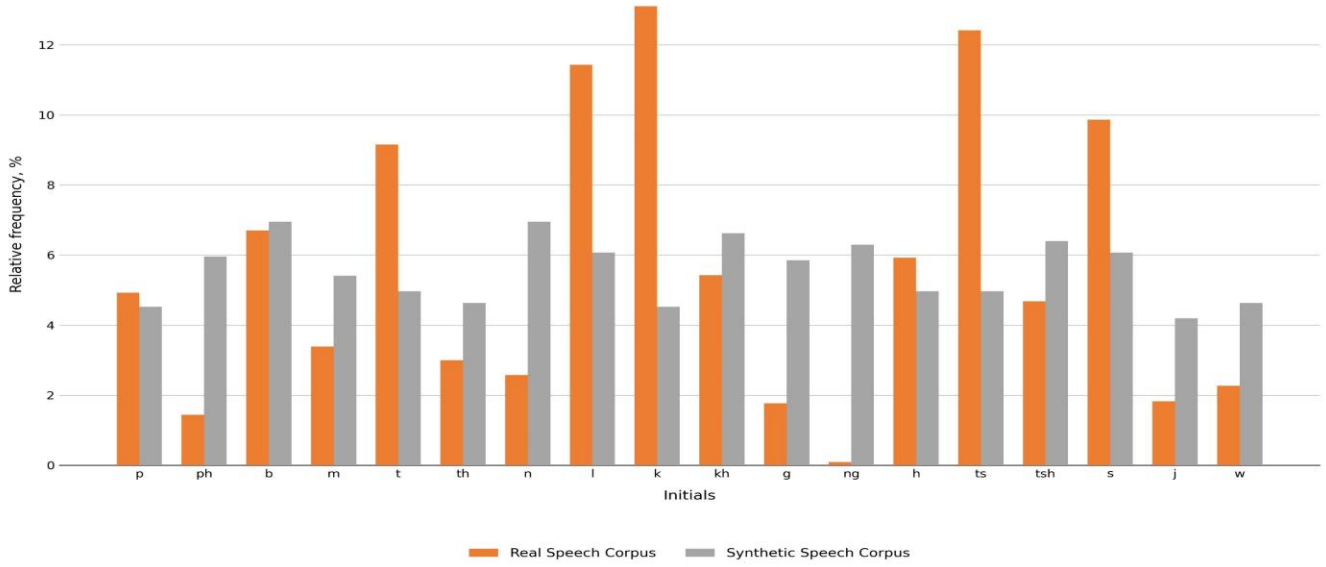

Fig. 2. Initials phoneme distribution.

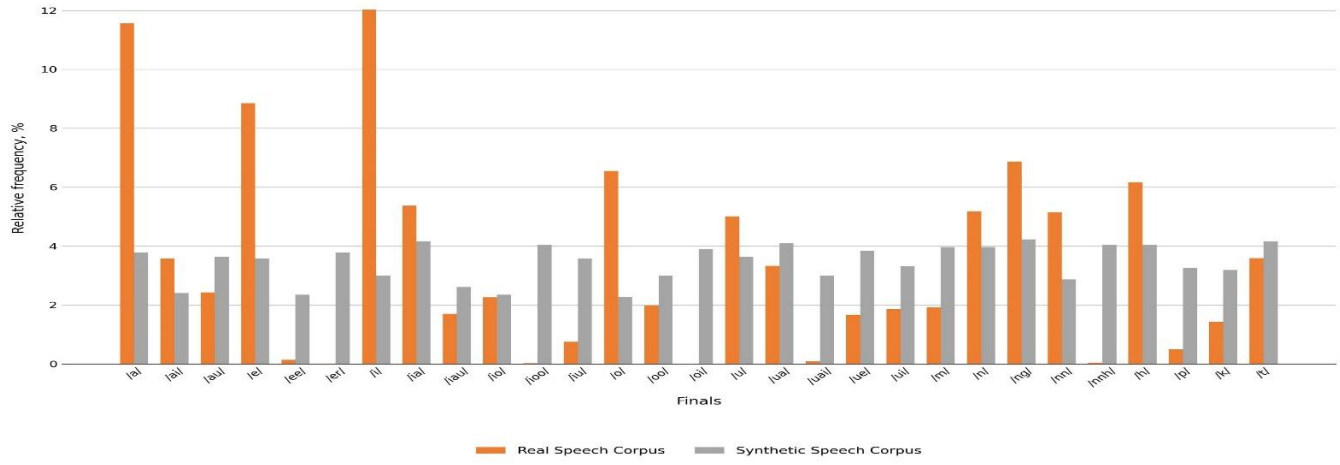

Fig. 3. Finals phoneme distribution.

C. Cross-Fading: Concatenative Artifact Mitigation

Concatenated speech may introduce artifacts at the boundaries between syllables, affecting the overall naturalness and fluidity of the synthesized output. These artifacts arise due to acoustic discontinuities between concatenated units, leading to unnatural transitions. To mitigate this issue, a cross-fading technique [30] was applied, which smooths transitions between syllables by gradually blending overlapping segments. This technique helped reduce abrupt changes and enhance the cohesion and flow of the synthesized speech.

The cross-faded audio signal $S(t)$ is defined as:

$$S(t) = \big(1 - \alpha(t)\big) \cdot A_n(t) + \alpha(t) \cdot A_{n+1}(t) \qquad (1)$$

For this study, the duration of each individual syllable unit signal $A_n(t)$ was fixed at 1.0 s, while the overlap duration T was set to 600 ms. The concatenation was achieved by applying a fade-out factor $(1 - \alpha(t))$ to the preceding syllable segment $A_n(t)$ and a fade-in factor $\alpha(t)$ to the succeeding syllable segment $A_{n+1}(t)$. These factors vary linearly with time (t) across the overlap duration T. Specifically:

- $\alpha(t) = \frac{t}{T}$ increases from 0 to 1 as t progresses through the overlap region, gradually bringing in the succeeding syllable segment.
- $1 - \alpha(t)$ decreases from 1 to 0 over the same time, fading out the preceding syllable segment.

The resulting cross-faded audio signal $S(t)$ smoothly transitioned between syllable units, ensuring natural-sounding audio continuity.

D. Numerical Tone Representation: Improve Tonal Learning

Penang Hokkien is a tonal language where variations in tone significantly alter word meaning. The model must correctly encode and distinguish these tonal variations for effective TTS training. The original Penang Hokkien read speech transcript uses diacritic characters (e.g., *"à"*, *"á"*, *"â"*) to indicate tones, but an alternative numerical tone representation (e.g., *"à"* → *"a3"*) was proposed in this study.

Using numerical tone markers in the input text instead of the original diacritic characters provides several advantages for low-resource TTS. First, the Penang Hokkien real speech corpus is limited to only 45 minutes, and individual diacritic characters may appear infrequently, making it difficult for the model to learn meaningful embeddings for these rare tokens due to limited exposure. In contrast, using numbers to represent tones consolidates tonal information into a smaller, shared set of tokens (e.g., *"à"*, *"è"*, and *"ì"* all map to the tone marker *"3"*). This significantly reduces token sparsity, allowing the model to learn tonal variations more effectively. Additionally, direct diacritic encoding would require 34 unique diacritic variations in the tokenizer, leading to a larger vocabulary size and increased embedding complexity. In comparison, numerical tone markers only require 7 additional tokens (1-7), minimizing vocabulary size and computational overhead. Since each token corresponds to a row in the model's embedding matrix, reducing the number of unique tokens simplifies the embedding space and enhances generalization, particularly in low-resource scenarios. Fig. 4 illustrates the mapping between diacritic tones and their corresponding numerical representations.

Diacritic tone representation :	a e i o u ng	á é í ó ú m̀ ńg	à è ì ò ù ǹg	ah eh ih oh uh	â ê î ô û m̂ n̂g	ā ē ī ō ū m̄ n̄g	a̍h e̍h i̍h o̍h u̍h m̍h n̍gh
Numerical tone representation :	1	2	3	4	5	6	7

Fig. 4. Diacritic tones in Penang Hokkien and their corresponding numerical tone representations.

E. Two-Stage Fine-Tuning: Phonetic Training and Prosodic Refinement

The Penang Hokkien TTS model was trained using a progressive two-stage fine-tuning approach designed to balance broad phonetic robustness with authentic prosodic naturalness. The model was initialized with a multitask-pretrained SpeechT5_TTS checkpoint trained on the high-resource English LibriTTS corpus [3], enabling cross-lingual knowledge transfer to the low-resource Penang Hokkien dialect.

The primary objective of the first stage was to establish a dense phonetic foundation and resolve the model's exposure to the comprehensive inventory of approximately 2,000 unique syllable-tone combinations. Based on our ablation study regarding optimal augmentation volume, the model underwent intermediate fine-tuning on 2 hours of the synthetic speech corpus. This established a 2:1 synthetic-to-real data ratio, which was identified as the inflection point for maximum performance. Before training, the tokenizer was extended to include numerical tone markers (1–7) to prevent the loss of critical tonal information. To accommodate these new tokens, the embedding layer was resized. During this stage, all layers were unfrozen to allow the model to fully integrate the newly established tonal relationships and adapt its hidden representation space to the modified input.

The second stage focused on refining acoustic modeling and speech feature generation while preserving the phonetic structures learned in Stage 1. The model was fine-tuned on the 45-minute real speech corpus to supply the natural prosody and authentic accentual nuances inherent in human recordings. To retain the learned phonetic structures, the token embedding layers, lower transformer layers, and self-attention mechanisms were frozen. Fine-tuning was restricted to the acoustic modeling layers (encoder and decoder) and the decoder's pre-net and post-net layers to optimize spectral feature prediction.

To convert the predicted mel-spectrograms into high-fidelity speech waveforms, we utilized a pretrained HiFi-GAN vocoder [3]. Unlike the acoustic model, the vocoder was not fine-tuned from scratch, as its architecture is designed to generalize fundamental speech characteristics, such as periodicity and loudness, across different languages and speakers [31].

Table III shows the hyperparameter settings used in model training.

TABLE III. MODEL TRAINING HYPERPARAMETERS

Hyperparameter	Value
learning_rate	1e-4
train_batch_size	4
eval_batch_size	2
seed	42
gradient_accumulation_steps	8
total_train_batch_size	32
optimizer	ADAMW_TORCH with β_1, β_2=(0.9,0.999) and ε =1e-8
lr_scheduler_type	linear
lr_scheduler_warmup_steps	500
training_steps	2000
mixed_precision_training	Native AMP

IV. EVALUATION

A. Objective Evaluation

The objective evaluation of synthesized Penang Hokkien speech employs on Mel-Cepstral Distortion (MCD) [32], a widely used spectral distance metric that quantifies the similarity between synthesized and groundtruth speech in terms of acoustic features. MCD measures the distance between the mel-frequency cepstral coefficients (MFCCs) extracted from both speech signals, providing an objective assessment of spectral fidelity. A lower MCD score indicates a smaller spectral difference, implying higher fidelity and better synthesis quality.

B. Subjective Evaluation

For subjective evaluation, Mean Opinion Score (MOS) [33] was conducted to assess the quality of the synthesized Penang Hokkien speech. MOS is a widely accepted perceptual metric that quantifies speech quality on a 5-point scale. To eliminate potential biases, the utterances were presented in a randomized order. A total of 10 native Penang Hokkien speakers participated in the evaluation. A structured questionnaire was utilized to evaluate the synthesized speech across three fundamental dimensions for tonal language synthesis [34]:

- Intelligibility: Measures how well a listener can understand the synthesized speech, including pronunciation accuracy, phoneme articulation, and word clarity.
- Naturalness: Evaluate how human-like the synthesized speech sounds, considering smoothness, voice quality, and absence of robotic artifacts.
- *Tone Accuracy* – Assesses how well the synthesized speech captures intonation, pitch variation, and stress patterns, which are essential in Penang Hokkien as a tonal language (Table IV).

TABLE IV. MEAN OPINION SCORE (MOS) EVALUATION SCALE

Score	Intelligibility	Naturalness	Tone Accuracy
1	No meaning understood	Unnatural, highly robotic	Not accurate, very big tonal shift
2	Effort required to understand	Inadequately natural	Inadequately accurate
3	Moderate effort required	Adequately natural	Adequately accurate
4	No major effort, mostly clear	Near natural	Near accurate, slight tonal shifts
5	No effort required, perfectly clear	Natural human speech	Very accurate, correct tone

V. RESULTS AND DISCUSSION

The primary contribution of this study is the empirical validation that, for tonal dialects like Penang Hokkien, addressing the "phonological gap" through a comprehensive phonetic inventory is more vital for synthesis quality than simply increasing the amount of training data through acoustic perturbations. While common data augmentation techniques like pitch-shifting [35] improve model robustness by simulating prosodic diversity, they do not introduce new phonetic information.

We conducted a controlled comparison between two distinct augmentation techniques, each providing 2 hours of synthetic speech data for intermediate fine-tuning before fine-tuning on the 45-min real speech data, with numerical input as tone representation.

1) Syllable-level concatenation: This augmentation technique constructed synthetic speech to cover the comprehensive set of approximately 2,000 Penang Hokkien syllables, ensuring every tonal and phonetic unit was represented during training.

2) Pitch-shifting: Augmentation was performed on the original single-speaker real speech corpus using the SoX tool [36] to simulate prosodic variation. Pitch was shifted within a narrow range of -0.5 to +0.5 semitones (step size: 0.5), and speed was perturbed by ± 10% (ratios: 0.9 and 1.1). This restricted range was selected to maintain speaker similarity, as larger perturbations are known to degrade perceived voice identity and naturalness [37][38]. While this augmentation technique provided an equivalent 2 hours of synthetic speech data, it remained restricted to the 1,300 syllables present in the original real speech corpus (Table V).

TABLE V. SYLLABLE-LEVEL CONCATENATION VS. PITCH-SHIFTING (PHONETIC COVERAGE VS. PROSODIC VARIATION)

Augmentation Technique	Unique Syllable Count	MOS ↑				MCD ↓
		Intelligibility	Naturalness	Tone Accuracy	Overall	
2 hr Syllable-Level Concatenation (Proposed)	2045	**4.12 ± 0.43**	3.95 ± 0.24	**3.69 ± 0.52**	**3.92 ± 0.40**	**20.07 ± 2.09**
2 hr Pitch-Shifting	1284	3.67 ± 0.26	**4.09 ± 0.17**	3.34 ± 0.51	3.70 ± 0.31	20.54 ± 1.86

The results reveal that the proposed syllable-level concatenation (MOS 3.92) consistently outperformed pitch-shifting (MOS 3.70). Pitch-shifting provides some benefits by introducing prosodic variation, which has been shown in prior work to improve model robustness [37]. However, this technique remains limited because it did not improve phonetic coverage, many syllables remain unseen during training. In contrast, the proposed syllable-level concatenation technique directly addressed phonetic coverage limitations. By systematically constructing synthetic speech to cover the comprehensive set of ~2,000 Penang Hokkien syllables, the model is exposed to every tonal and phonetic unit before final fine-tuning on a real speech corpus. This broader coverage is particularly critical for tonal languages, where accurate tone realization is tightly coupled with intelligibility [39].

Although pitch-shifting yields a slight improvement in perceived naturalness, native speaker evaluators consistently preferred the model intermediately fine-tuned on syllable-level concatenative synthetic speech for its clarity and tonal correctness. The higher ratings in intelligibility and tone accuracy indicate that in a low-resource tonal language setting, comprehensive phonetic coverage contributes more directly to synthesis quality than increased prosodic variation alone.

The superiority of the syllable-level concatenation technique is most evident when synthesizing Out-of-Vocabulary (OOV) syllables, such as *"gēr"*, *"lēeng"*, and *"ngāng"* in the sentence *"tsin gēr-lam khuànn lú tī tsia lēeng-ngāng-lēeng-ngang. (It's frustrating seeing you lengang-lengang here—just doing nothing.)"*. The effectiveness of this technique is visually substantiated through mel spectrogram analysis, which represents the spectral properties and energy distribution of synthesized speech. In Fig. 5, we compare the acoustic outputs for the OOV sentence across different training configurations.

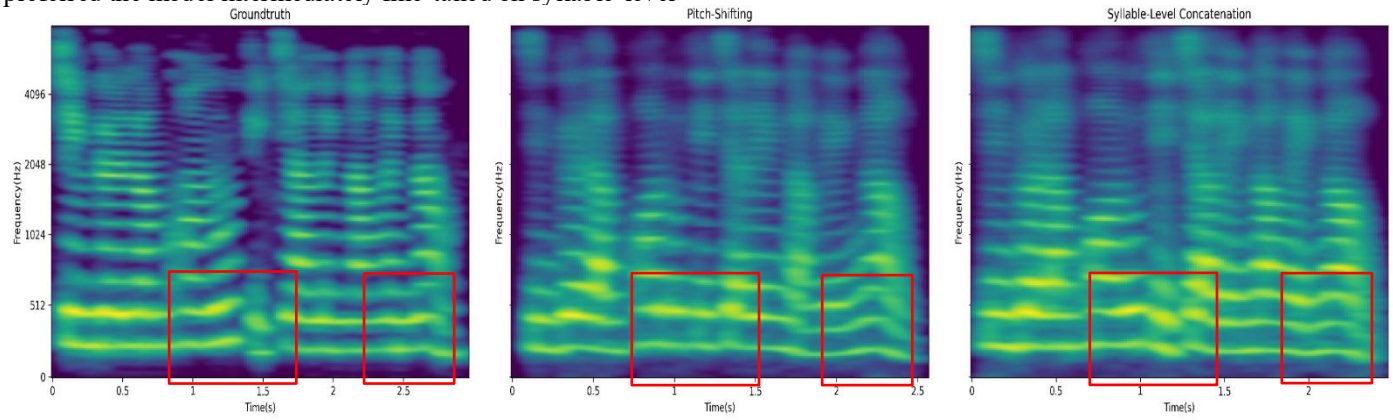

Fig. 5. Mel spectrogram analysis of OOV sentence.

Mel spectrograms represent the spectral properties of speech, where clear and well-defined harmonic structures (the bright horizontal bands) correlate with higher fidelity and intelligibility. The groundtruth spectrogram displays sharp, distinct harmonics, representing clear, natural speech. These features are indicative of high-fidelity natural speech with stable pitch contours. The output from the pitch-shifting model is noticeably blurrier, indicating weak harmonic definition and degraded spectral fidelity. While pitch-shifting introduces prosodic variety, it fails to equip the model with the necessary phonetic information to accurately reconstruct the spectral envelopes of OOV syllables. In contrast, the model trained with syllable-level augmentation produces significantly sharper spectrograms that closely resemble the groundtruth reference speech. The harmonic bands are more pronounced and stable, demonstrating that intermediate fine-tuning on a comprehensive phonetic inventory allows the model to learn correct phonetic and tonal structures for OOV syllables with high accuracy.

While the improved performance is consistent with the hypothesis that broader phonetic coverage benefits low-resource tonal dialect TTS, the observed gains are likely influenced by multiple interacting factors. First, the two-stage fine-tuning strategy may introduce curriculum learning effects, where intermediate training on synthetic speech establishes stable phonetic representations before refinement on natural speech. Second, the synthetic corpus produces denser token distributions by repeatedly exposing the model to shared syllable-tone patterns, potentially improving embedding stability for rare units. Third, intermediate pretraining on synthetic speech may also provide a regularization effect by exposing the model to a wider range of phonetic combinations prior to adaptation on the limited real speech corpus. Therefore, while the experiments suggest that phonetic completeness plays an important role, the individual contributions of these factors cannot yet be fully disentangled within the current study design.

VI. Ablation Studies

A. Cross-Fading on Concatenative Artifact Mitigation

During the preparation of the synthetic speech corpus, individual syllable units were recorded and concatenated to form short utterances. However, naive concatenation often results in audible discontinuities at unit boundaries, such as clicks, spectral jumps, and unnatural transitions, which can negatively impact downstream TTS training. To ensure the synthetic corpus provides a clean phonetic foundation, we evaluated the effectiveness of cross-fading technique in smoothing the acoustic transitions within the syllable-level concatenative synthetic speech.

We compared direct concatenation (0 ms) against a 600-ms cross-fade across various phonotactic contexts. To objectively quantify join quality, we utilized boundary-level Mel-Cepstral Distortion (MCD) [40] to measure spectral mismatch at the transition points (as primary source of perceptual artifacts).

TABLE VI. ACOUSTIC AND PERCEPTUAL EFFECTS OF CROSS-FADING

Condition	MOS (Naturalness) ↑	Boundary MCD (dB)↓
No cross-fading	3.07 ± 0.44	6.41 ± 0.92
600ms cross-fading	3.68 ± 0.38	3.74 ± 0.51

As shown in Table VI, the 600-ms cross-fading reduced boundary MCD from 6.41 dB to 3.74 dB, an improvement of approximately 41% compared to direct concatenation (no cross-fading). High MCD values in the no-cross-fading condition reflect abrupt spectral jumps, a well-documented artifact in concatenative synthesis. Applying a 600-ms cross-fading mitigated these jumps by gradually interpolating between spectral envelopes, yielding smoother formant trajectories and suppressing spurious high-frequency energy at the join. This smoothing also supports more natural F0 continuity across syllables. This reduction in spectral discontinuity translated to significant perceptual gains, with MOS ratings increasing from 3.07 (unnatural) to 3.68 (moderately natural). Evaluators reported a marked reduction in audible clicks and smoother transitions. These results confirm that cross-fading is an essential preprocessing step to ensure the model learns acoustic transitions consistent with natural coarticulation. Consequently, the 600-ms configuration was adopted for the final synthetic corpus construction.

B. Numerical vs. Diacritic Tone Representation Markers

In tonal dialects like Penang Hokkien, the tone markers used to represent pitch variations in the input text is a critical design factor that directly influences the model's ability to learn phonetic patterns. While diacritics are linguistically traditional, they often introduce high token sparsity and vocabulary complexity in low-resource scenarios, where data is insufficient to train rare character embeddings. We conducted this study to determine if numerical markers could simplify the embedding space and accelerate model's convergence without compromising tonal information.

The input text with diacritic tone markers (e.g. "à") and numerical tone markers (e.g. "a3") was compared across four dimensions:

1) Vocabulary coverage: The use of diacritic tone markers resulted in a larger and sparser token vocabulary, as each of the 34 unique diacritic variant was treated as a distinct token. Within the 45-minute real Penang Hokkien speech corpus, many of these diacritic variants appear only a few times, producing a long tail of low-frequency tokens. Numerical tone markers collapsed all these diacritic variants into a small, shared set of tone markers (1-7), significantly reducing tokenizer vocabulary by ≈35%.

2) Model convergence speed: Reusing a small set of numerical tone markers created a denser token distribution, allowing the model to learn tone embeddings earlier and with greater stability. Diacritics, conversely, resulted in many low-frequency tokens that produced noisy gradients and slowed down the optimization of embeddings. Since numerical tone markers appeared across many syllables, the model was exposed to more syllable-tone combinations for each tone class, enabling more consistent mappings between tone markers and intonation. Quantitatively, the numerical model consistently reached a stable validation loss plateau (0.5936) at Step 950 (Epoch≈39), compared to (0.5939) at Step 1150 (Epoch≈48) for the diacritis baseline, representing a faster convergence speed of ≈18%.

3) Synthesis tonal accuracy: Tonal accuracy was assessed by analyzing synthesis output. Tone substitution error was identified whenever the synthesized tone class did not match the expected tone. The numerical model achieved a significant reduction in error rates (6.5%) compared to diacritic model (18.2%).

4) Perceptual speech quality (XAB Listening Test) [41]: A preference test involving native speakers revealed a clear preference for the numerical model. 72.45% of responses favored the numerical-tone outputs due to more stable tonal contours and higher intelligibility, while 0% favored the diacritic model. The remaining 27.55% of responses indicated no preference, representing instances where perceptual differences were negligible.

These results are summarized in Table VII. These findings confirm that numerical tone markers are superior to diacritic tone markers for Penang Hokkien TTS model training, effectively improving both computational efficiency and human perceptual preference.

TABLE VII. NUMERICAL VS. DIACRITIC TONE MARKERS

Evaluation Dimension	Numerical Tone Markers	Diacritic Tone Markers	Outcome
Vocabulary coverage	Reduced by ~35% consolidated shared tokens	Larger, sparse token set due to infrequent diacritic forms	Numerical better (less sparsity)
Model convergence speed	~18% fewer steps to converge	Slower, converged in more training steps	Numerical faster
Tone mispronunciation rate	6.5% error rate	18.2% error rate	Numerical more accurate
Perceptual preference (XAB test)	72.45% preference (27.55% no preference)	0% preference	Numerical strongly preferred

C. Optimal Volume of Syllable-Level Concatenative Synthetic Speech Data

A critical challenge in low-resource TTS is identifying the "inflection point" where synthetic data provides maximum phonetic stability without introducing robotic artifacts or overfitting to mechanical prosody of concatenated units. We investigated this by fine-tuning four configurations with varying durations of the synthetic speech corpus (0-4 hours) followed by a constant 45-minute real speech corpus.

Among all evaluated configurations, Model III (2hr synthetic + 45min real) achieved the best overall performance, recording the highest MOS (3.92) and lowest MCD (20.07), indicating the closest spectral match to groundtruth reference speech. These results suggest that supplementing limited real speech data with syllable-level concatenative synthetic speech data helps the TTS model to learn better pronunciation patterns and tonal variations, yielding clearer and more accurate speech synthesis while maintaining naturalness (Tables VIII and IX).

TABLE VIII. TTS TRAINING DATA CONFIGURATIONS

Model	Synthetic Speech Data	Real Speech Data	Number of utterances	Total Syllable Count	Unique Syllable Count	Input Tone Representation
Model I	-	45 min	902	10507	1284	Numerical
Model II	1 hr	45 min	2110	13472	2045	Numerical
Model III	2 hr	45 min	3302	26502	2045	Numerical
Model IV	4 hr	45 min	5779	53481	2045	Numerical

TABLE IX. OBJECTIVE AND SUBJECTIVE EVALUATION OF SYNTHESIZED SPEECH

Model	MOS ↑				MCD ↓
	Intelligibility	Naturalness	Tone Accuracy	Overall	
Model I	3.71 ± 0.67	3.62 ± 0.42	3.38 ± 0.63	3.57 ± 0.57	21.04 ± 2.45
Model II	3.89 ± 0.39	3.78 ± 0.37	3.55 ± 0.44	3.74 ± 0.40	20.21 ± 1.97
Model III	**4.12 ± 0.43**	**3.95 ± 0.24**	3.69 ± 0.52	**3.92 ± 0.40**	**20.07 ± 2.09**
Model IV	3.92 ± 0.31	3.53 ± 0.33	**3.73 ± 0.45**	3.73 ± 0.36	20.14 ± 2.11

However, increasing the amount of synthetic speech data beyond an approximately 2:1 ratio did not yield meaningful improvements. Model IV (4hr synthetic + 45min real) showed only a marginal increase in tone accuracy (MOS 3.73 vs 3.69), but experienced slight declines in intelligibility and naturalness, and a higher MCD score of 20.14 ± 2.11. These results indicate a performance plateau, with signs of degradation beyond the optimal augmentation point. Several factors likely contributed to this degradation. First, syllable-level concatenative synthetic speech may introduce unnatural prosody and boundary artifacts. With more synthetic speech data, the model might overfit to these unnatural patterns, negatively affecting speech smoothness. Second, the phonetic diversity of synthetic speech data tends to diminish after a certain threshold; beyond two hours, additional examples primarily reinforce existing patterns rather than introducing new ones, raising the risk of overfitting to synthetic pronunciations. Third, the domain gap between synthetic and real speech may become more prominent as synthetic speech data dominates the training phase, reducing the effectiveness of real speech data fine-tuning in correcting artifacts.

Taken together, these findings suggest that an approximately 2:1 synthetic-to-real ratio (as implemented in Model III) achieved the optimal balance, maximizing tonal clarity and phonetic robustness while preserving natural human prosody.

VII. STATISTICAL VALIDATION

A. Inter-Rater Reliability

While a pool of 10 native speaker evaluators represents a compact sample size, recruiting qualified fluent speakers of an endangered, unwritten dialect presents acute resource constraints. To ensure the statistical reliability of this sample, the Kappa statistic [42] was utilized to measure the degree of agreement among the 10 native speaker evaluators (inter-rater agreement). Kappa values range from -1 to +1, where a score of +1 indicates perfect agreement and 0 indicates agreement purely by chance. The inter-rater agreement is calculated using the following formula:

$$K = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

where, P_o (observed agreement) represents the proportion of instances where evaluators assigned the same rating to a speech sample, while P_e (agreement expected by chance) accounts for the likelihood that evaluators would agree on the rating for a speech sample purely by chance.

The 10 native speaker evaluators participated in a calibration phase, where they rated each other's Penang Hokkien speech on a 5-point scale (1 = poor, 5 = excellent). The resulting Kappa score of 0.71 indicated substantial agreement based on the established interpretation scale for inter-rater reliability [43]. This high degree of consensus

confirmed that the evaluators shared a consistent and reliable standard for linguistic authenticity in Penang Hokkien, ensuring that the Mean Opinion Score (MOS) ratings provided in the subjective evaluation are both stable and reliable.

B. Statistical Significance of MOS Differences

To determine whether differences in perceived speech quality among all evaluated systems were statistically meaningful, a one-way ANOVA [44] was conducted across the MOS scores for all model configurations. The ANOVA revealed a significant main effect ($p < 0.05$), indicating that at least one system differed significantly from the others.

Post-hoc pairwise comparisons (Tukey's HSD [45]) were then performed to identify which specific model pairs exhibited statistically significant differences. Among the significant comparisons, Model III (MOS = 3.92) demonstrated a statistically significant improvement over the Baseline Model I trained solely on the 45-minute real dataset (MOS = 3.57), with $p < 0.05$. Significant differences were also observed between Model III and several other lower-performing augmentation variants, confirming the robustness of the proposed method across multiple comparative conditions.

VIII. Limitation

Despite the successful outcomes, this research has several limitations. The primary limitation is the use of a single-speaker, 45-minute real-speech dataset. This constraint reflects practical realities, limited funding, limited speaker availability, and the absence of any larger curated Penang Hokkien corpora. As such, the dataset used here represents one of the most complete resources available for this critically low-resource dialect. The motivation for proceeding with such a small corpus was to conduct a feasibility study demonstrating what is realistically achievable for endangered or under-documented languages under true data scarcity. Nevertheless, the limited dataset constrains the model's ability to generalize across diverse prosodic patterns and emotional expressions, and this gap contributes to the remaining MOS deficit relative to groundtruth speech. Second, standard automated metrics such as Word Error Rate (WER) or Character Error Rate (CER) via a standardized Automatic Speech Recognition (ASR) model were not viable due to the non-existence of a baseline ASR system for this dialect. Consequently, spectral fidelity via MCD and perceptual tracking via multi-dimensional MOS were prioritized.

IX. Conclusion

This study successfully trained the first Text-to-Speech model for Penang Hokkien, addressing the critical challenges of phonological sparsity and resource scarcity inherent in unwritten tonal dialects. By implementing a two-stage fine-tuning approach using the SpeechT5_TTS model architecture, it was demonstrated that a phonetically-rich foundation, constructed from a syllable-level concatenative synthetic corpus covering comprehensive syllable-tone combinations, is essential for achieving high-fidelity synthesis and tonal accuracy. Technical refinements, including a 600-ms cross-fading technique to mitigate boundary artifacts and the use of numerical tone markers to reduce token sparsity, contributed to an overall Mean Opinion Score (MOS) of 3.92. Further cross-language validation is required to definitively generalize these findings to other tonal families. We position this work as an initial case study and a technological blueprint for the digital preservation of endangered, unwritten tonal dialects.

The synthesized speech samples can be found at https://cclia19.github.io/penanghokkientts/.

Acknowledgment

This study was supported by the UTAR Research Fund (IPSR/RMC/UTARRF/2022-C1/J01), Universiti Tunku Rahman, Malaysia.

References

[1] H. T. Ling, "Penang Hokkien Speech Analysis," Medium. 2021. [Online] Available: https://linghuiting.medium.com/penang-hokkien-speech-analysis-329f20248bb.

[2] P. Nambiar, "Hokkien on its last legs, warns linguist", Free Malaysia Today (FMT), 2020. [Online] Available: https://www.freemalaysiatoday.com/category/leisure/2020/08/04/hokkien-on-its-last-legs-warns-linguist/ (Accessed: 19 Jan 2024)

[3] J. Ao, et al., "SpeechT5: Unified-Modal Encoder-Decoder Pre-training for Spoken Language Processing", arXiv preprint arXiv:2110.07205, 2021.

[4] H. Zen, et al., "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech", in Interspeech, Graz, Austria, 2019, pp. 1526-1530.

[5] F. S. Oliveira, et al., "CML-TTS: A Multilingual Dataset for Speech Synthesis in Low-Resource Languages", in International Conference on Text, Speech, and Dialogue, Cham: Springer Nature Switzerland, 2023. pp. 188-199.

[6] A. W. Black, "Cmu wilderness multilingual speech dataset", in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 5971-5975.

[7] J. Meyer, et al., "BibleTTS: a large, highfidelity, multilingual, and uniquely African speech corpus", in Interspeech, Incheon, Korea, 2022. pp. 2383-2387.

[8] A. Debnath, et al., "Low-Resource End-to-end Sanskrit TTS using Tacotron2, WaveGlow and Transfer Learning", in 2020 IEEE 17th India Council International Conference (INDICON), NSUT, Delhi, India, 2020, pp. 1-5.

[9] E. Cooper, "Text-to-speech synthesis using found data for low-resource languages", Ph.D dissertation, Columbia University, New York, US, 2019.

[10] A. Amalas, et al., "A multilingual training strategy for low resource Text to Speech", arXiv preprint arXiv:2409.01217, 2024.

[11] M. Baali, et al., "Unsupervised data selection for TTS: using Arabic Broadcast News as a case study", arXiv preprint arXiv:2301.09099, 2023.

[12] W. Fang, Y. A. Chung, and J. Glass, "Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models", arXiv preprint arXiv:1906.07307, 2019.

[13] Y. Jia, et al., "PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS", in Interspeech, Brno, Czechia, 2021, pp. 151-155.

[14] A. H. Liu, "Towards unsupervised speech recognition and synthesis with quantized speech representation learning", in ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7259-7263.

[15] H. Zhang and Y. Lin, "Unsupervised Learning For Sequence-to-sequence Text-to-speech For Lowresource Languages", in Interspeech, Shanghai, China 2020, pp. 3161-3165.

[16] K. Lakhotia, et al., "On generative spoken language modeling from raw audio", Transactions of the Association for Computational Linguistics, vol. 9, 2021, pp. 1336-1354.

[17] J. Ni, et al., "Unsupervised Text-to-Speech Synthesis by Unsupervised Automatic Speech Recognition", in Interspeech, Incheon, Korea, 2022, pp. 461-465.

[18] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning", in 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 2017, pp. 301-308.

[19] Y. Ren, et al., "Almost unsupervised text to speech and automatic speech recognition", in International Conference on Machine Learning, Long Beach, CA, USA, 2019, pp. 5410-5419.

[20] Z. Byambadorj, et al., "Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation", EURASIP Journal on Audio, Speech, and Music Processing, vol. 2021, no. 1, 2021, pp. 1-20.

[21] K. K. Lakshminarayana, et al., "Low-Resource Text-to-Speech Synthesis Using Noise-Augmented Training of ForwardTacotron," in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1-5.

[22] T. Tao, et al., "End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning", in Interspeech, Graz, Austria, 2019. pp. 2075-2079.

[23] J. Xu, et al., "Lrspeech: Extremely low-resource speech synthesis and recognition", in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020. pp. 2802-2812.

[24] Y. Zhang, et al., "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning", in Interspeech, Graz, Austria, 2019, pp. 2080-2084.

[25] K. Azizah, M. Adriani, and W. Jatmiko, "Hierarchical Transfer Learning for Multilingual, Multi-Speaker, and Style Transfer DNN-Based TTS on Low-Resource Languages", IEEE Access, 2020, Vol. 8, pp. 179798-1798812.

[26] C. H. Tan, Penang Hokkien for Penangites and Tourists. MPH Group Publishing Sdn Bhd, 2014.

[27] Learn Penang Hokkien. [Facebook]. Available at: https://www.facebook.com/groups/learnpenanghokkien/

[28] Y. L. Lai, J. Y. M. Khaw, S. P. Lim, and T. P. Tan, "Text Normalisation of Penang Hokkien Dialect Leveraging Adapted Soundex Algorithm," 2024 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS), Bangkok, Thailand, pp. 48-54, Sep. 2024

[29] Y. Tabet, A. Boustil, and K. Baiche, "Speech Analysis-Synthesis using Deep Neural Networks: A Review", in The 1st National Workshop on Wireless Network, Cloud Computing and Cryptography (WWN3C'2023), Boumerdes, Algeria, 2023, pp. 110-114.

[30] A. Franck, "Efficient Algorithms and Structures for Fractional Delay Filtering Based on Lagrange Interpolation", Journal of the Audio Engineering Society, vol. 56, no. 12, 2009, pp. 1036-1056.

[31] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis", Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 17022-17033.

[32] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment", in Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, vol. 1, 1993, pp. 125–128.

[33] ITU-T Recommendation P.85, "A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices", International Telecommunication Union, 1994.

[34] T. T. T. Nguyen, "HMM-based Vietnamese Text-To-Speech : Prosodic Phrasing Modeling, Corpus Design System Design, and Evaluation," Ph.D. dissertation, Université Paris Sud-Paris XI; Institut Polytechnique (Hanoï), 2015.

[35] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 9, 2015, pp. 1469-1477.

[36] SoX: Sound eXchange audio manipulation tool. 2024. [Online] Available: http://sox.sourceforge.net.

[37] T. Ko, et al., "Audio augmentation for speech recognition," in Interspeech, Dresden, Germany, 2015, pp. 3587-3589.

[38] A. Polyak, et al., "Speech resynthesis from discrete disentangled self-supervised representations," in Interspeech, Brno, Czechia, 2021, pp. 3615-3619.

[39] J. Tang and M. Li, "End-to-End Mandarin Tone Classification with Short Term Context Information," in 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 2021, pp. 878-883.

[40] J. Vepa, S. King, and P. Taylor, "New Objective Distance Measures for Spectral Discontinuities in Concatenative Speech Synthesis," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, 2002, pp. 223-226.

[41] J. Zhong, et al., "Pairwise Evaluation of Accent Similarity in Speech Synthesis," *arXiv preprint arXiv:2505.14410*, 2025.

[42] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability", Educational and Psychological Measurement, vol. 33, no. 3, 1973, pp. 613–619.

[43] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," Biometrics, vol.33, no.1, 1977, pp. 159-174.

[44] Fisher, R. A., Statistical Methods for Research Workers. Oliver and Boyd, 1925.

[45] Tukey, J. W., "Comparing individual means in the analysis of variance," Biometrics, vol. 5, no. 2, pp. 99-114, Jun. 1949.