Optimizing Filipino Text-to-Speech Synthesis: Integration of Generative Models

Josephine Jane Gapuz*, Dean Kyle Mariano*, Jezler Recto*, Rhandley Cajote*, John Cairu Ramirez*
*Electrical and Electronics Engineering Institute
University of the Philippines Diliman
Quezon City, Philippines
{josephine.jane.gapuz, dean.kyle.mariano, jezler.recto, rhandley.cajote, john.cairu.ramirez}@eee.upd.edu.ph

***Abstract*—This paper explored the development of an optimized text-to-speech (TTS) system for Filipino by integrating two advanced generative models: FastSpeech 2 and Tacotron 2. Since Filipino is a low-resource language for TTS, we leveraged the Filipino Speech Corpus—which includes over 150 hours of annotated audio—to create model-specific datasets. Preprocessing steps using the Montreal Forced Aligner, phoneme-level pitch and energy normalization were implemented in FastSpeech2 to ensure precise spectral mapping. In contrast, Tacotron 2 was trained on gender-specific datasets to better capture natural prosody variations. Each synthesis pipeline paired its model with a state-of-the-art vocoder: WaveGlow for Tacotron 2 and a universal HiFi-GAN for FastSpeech 2.**

Objective evaluation was performed using the mel-cepstral distance metric to compare the spectral characteristics of the synthesized speech against ground truth recordings. Subjective listening tests including Mean Opinion Score (MOS) assessments and AB tests were conducted in a controlled lab environment in which the insights were provided to analyze perceptual naturalness and intelligibility of synthesized speech. Overall, the results indicate that while both models contribute to improving Filipino TTS, the FastSpeech 2 surpassed Tacotron 2 in MCD however Tacotron 2 produced higher MOS ratings and more natural-sounding speech than FastSpeech 2.

This study offers a useful framework and insights in enhancing TTS systems despite using low-resource languages, which could help expand access to digital voice applications and guide future research in speech synthesis.

***Index Terms*—TTS, Tacotron 2, FastSpeech 2, Filipino speech synthesis, montreal forced aligner**

I. Introduction

Text-to-speech (TTS) systems convert written text into synthetic speech and have significantly contributed to the digital accessibility revolution [4]. The field of TTS is multidisciplinary, intersecting linguistics, signal processing, and deep learning—particularly generative models that aim to produce human-like prosody and articulation.

The development of natural and intelligible TTS systems for low-resource languages like Filipino presents unique challenges due to limited datasets. Among these, Tacotron 2 and FastSpeech 2 have emerged as widely adopted architectures for speech synthesis in multiple languages, demonstrating reliable performance in both high- and low-resource settings [1], [2]. Tacotron 2 is known for generating expressive and natural-sounding speech in various languages, while FastSpeech 2 is a more recent approach that improves training efficiency and synthesis speed [2]., making it a compelling choice for evaluating TTS in Filipino for the first time in this context.

Previous studies on Filipino TTS used Tacotron 2 and WaveGlow with promising results [1]. Moreover, some studies have also experimented with Tacotron 2-based pipelines. For instance, Tatoy et al. [5] introduced a harmonic-plus-noise network with linear prediction and perceptual weighting filters to enhance the Tacotron 2 baseline. Their system achieved significantly lower Mel-Cepstral Distortion (MCD) and higher perceptual quality through vocoder-level innovations. Meanwhile, earlier work by Renovalles et al. [1] compared unit selection and deep learning systems using Tacotron 2, but noted low MOS scores due to limited training steps and lack of data augmentation.

For FastSpeech 2, forced alignment systems like the Montreal Forced Aligner (MFA) also helped in determining the exact timing of each phoneme in a spoken word, which is critical for generating natural prosody and accurate phoneme-to-frame alignment [3].

This project designed a TTS system specifically for the Filipino language, incorporating generative models in generating human-like speech with a natural prosody. To evaluate the quality of synthesized speech, both subjective and objective evaluation methods were used, such as the Mean Opinion Score (MOS) assessment, AB listening test, and Mel-Cepstral Distortion (MCD) measure.

Overall, this project aimed to further enhance communication accessibility for the local community through the integration of Filipino speech corpora, acoustic models, and the use of deep learning-based text-to-speech models.‘

II. Methodology

A. Data Preparation

The data preparation process for the FastSpeech 2 and Tacotron 2 models was conducted in three main stages: data acquisition, preprocessing, and alignment.

1) Data Acquisition: This study utilized the Filipino Speech Corpus (FSC), created by the University of the Philippines Diliman Electrical and Electronics Engineering Institute Digital Signal Processing Laboratory. The FSC comprises four volumes of speech recordings from 115 speakers, totaling over 150 hours of annotated audio.

For this project, two distinct speaker subsets were tailored to the architectural needs of each TTS model. FastSpeech 2 was trained using 37 speakers (18 male, 19 female), while Tacotron 2 used 18 speakers (7 male, 11 female). The selection criteria prioritized recordings with clear articulation, consistent prosody, minimal background noise, and accurate transcriptions to ensure stable training. Tacotron 2’s autoregressive

979-8-3315-6946-4/25/$31.00 ©2025 IEEE

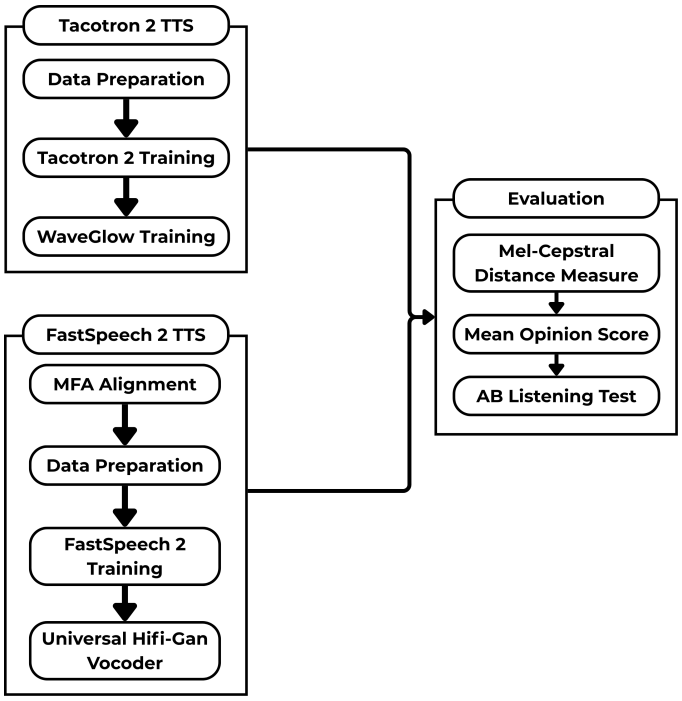

Fig. 1. Methodology Overview

structure is more sensitive to noisy or varied data, so it was trained on smaller, curated gender-specific datasets to reduce variability. In contrast, FastSpeech 2 benefits from broader speaker diversity to improve multi-speaker generalization and speaker embedding robustness. While the subsets differ in size and structure, there is partial overlap, 11 speakers are shared across both models (e.g., IDs 002, 003, 015, 025, 070, 081, 098), allowing for a fair yet context-aware comparison. This design balances architectural strengths with data quality constraints, enabling both models to be trained under conditions that suit their intended performance goals.

2) Data Preprocessing: The recordings and corresponding transcriptions were segmented into durations of less than 10 seconds to mimic other datasets, LJSpeech or LibriTTS, used in TTS. Shorter segments help ensure more stable alignment and improves training efficiency. For FastSpeech 2, additional preprocessing steps were necessary, including dictionary augmentation and acoustic model training. Dictionary augmentation was carried out using a grapheme-to-phoneme (G2P) model to generate phonetic transcriptions for out-of-vocabulary words [7]. An acoustic model was then trained using the segmented corpus and the augmented lexicon to enable accurate phoneme-to-audio alignment [8].

3) Data Alignment for FastSpeech 2: Phoneme-level time alignments required for training FastSpeech 2 were obtained using the Montreal Forced Aligner (MFA), an open-source speech-text alignment system built upon the Kaldi framework [9].

Using the split corpus, augmented phonetic dictionary, and trained acoustic model, MFA generated time-aligned transcriptions in the form of TextGrid files. These files provided precise alignment of phonemes and words to the audio, forming structured and reliable inputs of the FastSpeech 2 training pipeline.

B. Training Tacotron 2 and WaveGlow Vocoder

1) Dataset Preprocessing: The Tacotron 2 training pipeline required additional preprocessing to conform to the dataset structure used in LJSpeech-1.1. Recordings from the 7 male speakers were consolidated into a single male dataset, while those from the 11 female speakers formed a separate female dataset. Accordingly, Tacotron 2 was trained on two distinct gender-specific datasets.

The resulting datasets consisted of audio samples in .wav format and an associated metadata.csv file. From this metadata, file list directories were generated, each containing text files that mapped audio file paths to their corresponding transcriptions. These lists facilitated the organization of the dataset into training, validation, and testing partitions. Transcriptions were maintained at the sentence level to ensure consistency with the LJSpeech-1.1 format.

Mel-spectrograms were extracted from the audio files and stored separately. These served as the model inputs during Tacotron 2 training. The specific preprocessing parameters used in this stage are summarized in Table I.

TABLE I
TACOTRON 2 PREPROCESSING PARAMETERS

Parameter	Value
Sampling rate	22050 Hz
Filter length	1024
Hop length	256
Window length	1024
Mel Fmin	0
Mel Fmax	8000
Mel channels	80

2) Tacotron 2 Model Training: The Tacotron 2 model was implemented based on the official NVIDIA Tacotron 2 repository [6]. A pre-trained model, originally trained on the LJSpeech-1.1 dataset for 6000 epochs with a batch size of 100, was fine-tuned using a subset of the Filipino Speech Corpus. Fine-tuning was conducted for an additional 700 epochs using an NVIDIA A100 GPU on Google Colab. The training process required approximately 16 hours for the male dataset and 28 hours for the female dataset. The training parameters used during fine-tuning are summarized in Table II.

TABLE II
TACOTRON 2 TRAINING PARAMETERS

Parameter	Value
Learning rate	1e-3
Epochs	700
Batch size	128
Weight decay	1e-6
Grad clip thresh	1.0
cuDNN	Enabled
AMP	Enabled

3) WaveGlow Vocoder Training: To synthesize waveforms from the outputs of the Tacotron 2 model, the WaveGlow vocoder from the same repository [6] was employed. The audio samples used for training were sourced form the wavs directory. A pre-trained WaveGlow model, originally trained

on LJSpeech-1.1 dataset for 3000 epochs with a batch size of 10, was fine-tuned using a subset of the FSC. Fine-tuning was conducted for an additional 500 epochs on an NVIDIA A100 GPU via Google Colab, requiring approximately 7.25 hours for the male dataset and 12.5 hours for the female dataset. The training parameters utilized are detailed in Table III.

TABLE III
WAVEGLOW TRAINING PARAMETERS

Parameter	Value
Epochs	500
Batch size	16
Segment length	8000
Weight decay	0
Grad clip thresh	65504.0
cuDNN	Enabled
cuDNN benchmark	Enabled
AMP	Enabled

C. Training FastSpeech 2 and Utilizing Universal HiFi-Gan Vocoder

1) Data Preprocessing: Following initial preprocessing, the subset for FastSpeech 2 underwent additional cleaning and feature extraction. The specific parameters used during this stage are listed in Table IV.

In addition to this, phoneme-level pitch and energy normalization were applied and four key acoustic features, namely, duration, energy, mel spectrogram, and pitch, were extracted for each utterance. The subset was then partitioned into training and validation sets, with the latter used to monitor and evaluate model performance throughout training.

TABLE IV
FASTSPEECH 2 PREPROCESSING PARAMETERS

Parameter	Value
Validation size	512
Sampling rate	22050 Hz
Max wav value	32768.0
Filter length	1024
Hop length	256
Window length	1024
Mel channels	80
Mel Fmin	0
Mel Fmax	8000 Hz (HiFi-GAN)

2) Speaker Encoding: To enable multi-speaker modeling, each of the 37 speakers in the dataset was assigned a unique identifier, which was used to generate speaker embeddings. Table V presents the mapping between speaker identifiers and corresponding speaker IDs.

3) Phoneme Adaptation: The FastSpeech 2 framework, as implemented in FastSpeech2 GitHub repository [11], was originally configured to support English and Chinese phonemes. To accommodate Filipino phonemes, 18 vowels and 25 consonants were incorporated as shown in Table VI.

4) Model Training: The model was trained for 200,000 steps on an NVIDIA A100 GPU using Google Colab. The total training time was approximately 12 hours. The hyperparameters during training are listed in Table VII.

TABLE V
SPEAKER ID MAPPING

Speaker	ID	Speaker	ID	Speaker	ID
025	0	026	1	103	2
081	3	063	4	038	5
065	6	098	7	002	8
013	9	008	10	106	11
076	12	041	13	012	14
101	15	109	16	009	17
033	18	080	19	075	20
003	21	031	22	082	23
099	24	107	25	007	26
070	27	032	28	036	29
039	30	062	31	108	32
068	33	097	34	028	35
015	36				

TABLE VI
PHONEME ADAPTATION

Vowels	Consonants
AA, AE, AH, AO, AW, AX, AY, EH, ER, EY, IH, IX, IY, OW, OX, OY, UH, UW	B, CH, D, DH, F, G, H, JH, K, L, M, N, NG, NY, P, R, S, SH, T, TH, V, W, XL, Y, Z

D. Inference and Evaluation

1) Tacotron 2 Inferencing: Inferencing was conducted using the Tacotron 2 and WaveGlow checkpoints (.pt file) produced during training, following the setup from the original repository [6]. Key parameters used for inference are presented in Table VIII.

To assess whether the models effectively learned over time, the final training and validation losses for Tacotron 2 and WaveGlow were recorded and are shown in Tables IX and X. Their differences provides insignt into potential overfitting or underfitting during the training process.

2) FastSpeech 2 Inferencing: A universal HiFi-GAN vocoder [12], pre-trained on LibriSpeech, VCTK, and LJSpeech datasets, was utilized for inference. The best checkpoint for inference was selected based on objective performance, specifically Mel-Cepstral Distortion (MCD) evalua-

TABLE VII
FASTSPEECH 2 TRAINING PARAMETERS

Parameter	Value
Batch size	32
Betas	(0.9, 0.98)
Epsilon	1e-9
Grad clip threshold	1.0
Warm-up steps	4000
Total training steps	200,000

TABLE VIII
TACOTRON 2 INFERENCING PARAMETERS

Parameter	Value
Include warm up	Enabled
FP16	Enabled
Sigma infer	0.7
Denoising strength	0.07

TABLE IX
TACOTRON 2 TRAINING LOSS AND VALIDATION LOSS

Gender	Epochs	Training Loss	Validation Loss	Difference
Male	700	0.1663	0.3549	0.1886
Female	700	0.1359	0.2505	0.1146

TABLE X
WAVEGLOW TRAINING LOSS AND VALIDATION LOSS

Gender	Epochs	Training Loss	Validation Loss	Difference
Male	500	-7.7449	-7.7353	0.0095
Female	500	-6.8449	-6.8089	0.0360

tions. Model performance was assessed at 10,000-step intervals to identify the optimal checkpoint, as summarized in Table XI.

Lower-step checkpoints exhibited improved objective and subjective performance, as evidenced by smaller training and validation loss differences and superior MCD scores. Loss curves revealed that training began to plateau around 30,000 steps, suggesting convergence. Notably, the checkpoint at 34,000 steps achieved the lowest MCD values of 4.07 for male and 5.51 for female speakers, indicating the highest spectral similarity to the original recordings.

TABLE XI
MODEL EVALUATION ACROSS TRAINING STEPS

Step	Train Loss	Val Loss	Diff.	MCD (Male/Female)
10k	1.9162	2.5743	0.6581	4.68 / 6.11
20k	1.3587	2.6886	1.3299	4.43 / 5.80
30k	1.2981	2.7167	1.4186	4.23 / 5.51
33k	1.3117	2.7475	1.4258	4.20 / 5.43
34k	1.3385	2.7237	1.3852	4.07 / 5.51
35k	1.3292	2.7336	1.4044	4.43 / 5.55
40k	1.1360	2.7645	1.6285	4.13 / 5.54
50k	1.2181	2.8003	1.5822	4.20 / 5.68

3) Mel-Cepstral Distance Metric: For the objective metric, the mel-cepstral distance metric was utilized to quantify spectral distortion between ground truth and synthesized speech for both models. The implementation provided in [10] was employed for this computation. This metric specifically assesses differences in the spectral envelopes of the two audio

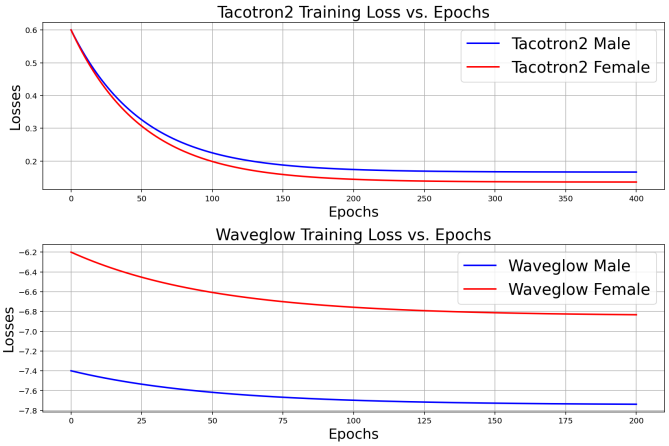

Fig. 2. Tacotron 2 and WaveGlow Training Loss

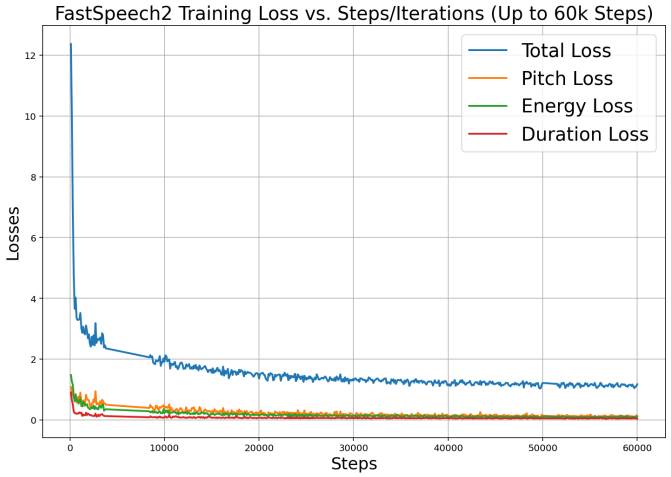

Fig. 3. FastSpeech 2 Training Loss

signals, offering an objective measurement of how closely the synthesized speech matches the original recordings.

4) Mean Opinion Score: A subjective listening test was conducted following the ITU-T P.800 recommendation for speech quality assessment [13]. The test took place in the Whisper Room of the Digital Signal Processing (DSP) laboratory at the EEE department, ensuring minimal external noise interference. Thirty participants rated 40 audio samples, consisting of 20 FastSpeech 2 (FS2) samples and 20 Tacotron 2 (TC2) samples, on a five-point scale. The audio samples for each model are grouped into female and male voices, with each gender having 5 samples from both custom input and ground truth synthesis.

Participants rated each sample based on the level of effort required to understand the speech:

- 1 - No meaning understood with any feasible effort
- 2 - Considerable effort required
- 3 - Moderate effort required
- 4 - Attention necessary; no appreciable effort required
- 5 - Complete relaxation possible; no effort required

Each participant provided subjective assessments on the effort required to understand the meaning of sentences.

5) AB Listening Test: To evaluate the perceptual quality of synthesized speech, we conducted an AB preference listening test comparing each synthesized system, FastSpeech 2 and Tacotron 2, against ground truth recordings. Participants were presented with pairs of audio samples, one from the ground truth (GT) and the other from the synthesized speech, in randomized order. They were then asked to choose which sample sounded the best to them. Thirty participants were involved in the test. Fifteen participants evaluated 20 FastSpeech 2 vs. ground truth pairs, while the other fifteen evaluated 20 Tacotron 2 vs. ground truth pairs.

To ensure fair comparison, white noise was injected into the ground truth recordings to match the noise level typically present in the synthesized outputs. White noise was generated in MATLAB based on the average noise levels detected in the synthesized samples, and added to the ground truth recordings at a signal-to-noise ratio (SNR) of 40 dB.

To assess the statistical significance of listener preferences, a binomial test was performed using jamovi [14]. This non-parametric statistical test is appropriate for analyzing binary outcome data under the null hypothesis that participants have no preference between two samples (0.5 probability of choosing either).

III. Results and Discussion

A. Mel-Cepstral Distance (MCD) Analysis

Lower Mel Cepstral Distortion (MCD) values indicate a closer resemblance to natural speech. In our experiments, FastSpeech 2 achieved lower MCD values compared to Tacotron 2. FastSpeech 2 produced an average MCD of 7.95 dB for male speakers and 9.18 dB for female speakers, while Tacotron 2 registered average values of 14.95 dB and 10.44 dB for male and female speakers, respectively. One factor influencing these results is the structure of the training dataset used for Tacotron 2. In the experiments, the multi-speaker dataset was reformatted into a single-speaker version. This reformatting likely diminished unique speaker-specific characteristics, such as consistent voice quality, which are vital for effective spectral modeling. In contrast, FastSpeech 2 was trained using the original multi-speaker structure, which allowed it to capture the natural variability among speakers.

Compared to recent work in Filipino speech synthesis, notably the study by Tatoy et al. [5], our MCD scores are higher in absolute value. Their Harmonic-plus-Noise (H+N) systems, such as HN-PWG and HN-PWG-PW, achieved significantly lower MCD values (e.g., 0.6983 and 0.7048, respectively), attributed to perceptual weighting and linear prediction techniques. However, our study utilizes mainstream neural TTS architectures without specialized enhancements, focusing on the integration of multispeaker corpora and standardized vocoders. Despite the higher distortion values, FastSpeech 2 still achieves a substantial improvement over our Tacotron 2 implementation and offers a competitive benchmark in Filipino TTS using a simpler pipeline.

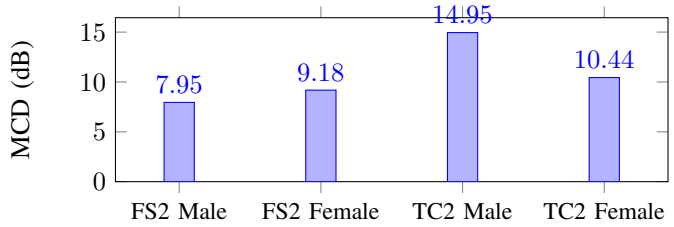

Fig. 4. Mel-Cepstral Distance Measure Comparison

B. MOS Listening Test Results

The MOS results clearly show a difference in perceived speech quality between FastSpeech 2 and Tacotron 2. Participant ratings indicate variations in naturalness, intelligibility, and the listening effort required to understand the synthesized speech. A gender-based analysis of the MOS scores reveals differences in listener perception among male and female groups. Overall, Tacotron 2 received an average MOS of 4.31, compared with 3.20 for FastSpeech 2, suggesting that listeners expended less effort when understanding the speech generated by Tacotron 2. In gender-specific evaluations, Tacotron 2 was rated 4.59 for female voices and 4.04 for male voices, whereas FastSpeech 2 scored 3.05 for female voices and 3.34 for male voices. These results indicate that the Tacotron 2 model produces synthesized speech that is subjectively superior and easier to comprehend.

These results align with prior works emphasizing Tacotron-based models' capacity to generate smoother and more natural prosody. In particular, Tatoy et al. [5] reported that their Tacotron 2-based system scored significantly lower, with a MOS of 2.31 ± 0.27, highlighting the perceptual advantage introduced by their H+N vocoding techniques. The current results suggest that Tacotron 2, combined with WaveGlow and dataset merging, can produce perceptually high-quality speech even without architectural modifications.

Historical comparison with the earlier work by Renovalles et al. [1] further contextualizes our improvements. Their Tacotron 2 system, built with limited data and trained for fewer steps, produced a maximum MOS of 2.01 after applying data augmentation techniques (TACO-DA). Even their unit selection system (MARY-B) only reached 3.05. In contrast, our Tacotron 2 implementation surpasses both benchmarks despite not using complex prosody modifications or voice conversion. These results support the effectiveness of our dataset structure and model setup in improving perceptual quality under similar resource constraints.

TABLE XII
Mean Opinion Score

Category	FastSpeech 2	Tacotron 2
Overall	3.20	4.31
Female	3.05	4.59
Male	3.34	4.04

C. AB Listening Test Results

In the AB listening test, listeners showed a strong preference for natural ground-truth speech over the synthesized output from both the FastSpeech 2 and Tacotron 2 models. These results confirm that listeners can reliably distinguish between natural and synthesized speech. In particular, the perceptual gap was greater in FastSpeech 2 than in Tacotron 2. Participants selected ground truth speech in 294 out of 300 comparisons for FastSpeech 2 and in 240 out of 300 comparisons for Tacotron 2. These results indicate that listeners consistently distinguished natural recordings from synthesized speech.

TABLE XIII
AB Listening Test

Category	Ground Truth	Synthesized
FastSpeech 2	294	6
Tacotron 2	240	60

A binomial test was conducted to assess the null hypothesis of no preference (i.e., a 50/50 chance). The test yielded a p-value of less than 0.001 for both models. The 95% confidence interval for the proportion of ground truth selections ranged from 0.957 to 0.993 for FastSpeech 2 and from 0.750 to 0.844 for Tacotron 2. These findings reinforce that listeners overwhelmingly preferred natural speech over the synthesized

output of both models. The results are better shown in Fig. 5 and Fig. 6

These findings reaffirm the MOS results: Tacotron 2 generates more perceptually convincing speech than FastSpeech 2. While Tatoy et al. [[5] did not use AB tests, their extensive MOS and statistical evaluations (e.g., ANOVA and Tukey-Kramer tests) strongly support the superior listener experience of HN-enhanced architectures. Our AB results, while expectedly favoring natural speech, demonstrate that Tacotron 2 narrows the perceptual gap more effectively than FastSpeech 2 in our test conditions. This complements Renovalles et al.'s [1] findings, where earlier Tacotron 2 models exhibited significant perceptual limitations—now overcome in our more recent configurations.

Proportion Test (2 Outcomes)

Binomial Test

						95% Confidence Interval	
	Level	Count	Total	Proportion	p	Lower	Upper
FastSpeech 2	Ground Truth	294	300	0.980	< .001	0.95698	0.9926
	Synthesized	6	300	0.020	< .001	0.00737	0.0430

Note. H_a is proportion ≠ 0.5

Fig. 5. Binomial Test result for FastSpeech 2

Proportion Test (2 Outcomes)

Binomial Test

						95% Confidence Interval	
	Level	Count	Total	Proportion	p	Lower	Upper
Tacotron 2	Synthesized	60	300	0.200	< .001	0.156	0.250
	Ground Truth	240	300	0.800	< .001	0.750	0.844

Note. H_a is proportion ≠ 0.5

Fig. 6. Binomial Test result for Tacotron 2

IV. Conclusion and Recommendations

In conclusion, this study explored the performance of two generative TTS models: Tacotron 2 and FastSpeech 2 for synthesizing Filipino speech. A key difference lies in their training approach: Tacotron 2 was fine-tuned from a pretrained English model, which helped it quickly adapt to the Filipino dataset despite limited data. This fine-tuning process allowed the model to retain learned prosodic and linguistic patterns, resulting in more natural and expressive speech output. In contrast, FastSpeech 2 was trained from scratch using a multi-speaker Filipino dataset, allowing it to generate various voices but also making it more dependent on data quality and speaker variation.

These methodological differences significantly influenced the evaluation results. Tacotron 2 achieved higher subjective scores such as a Mean Opinion Score (MOS) of 4.31 and stronger AB test preference—due to its smoother prosody and naturalness. However, FastSpeech 2 demonstrated better objective performance, with lower Mel-Cepstral Distance (MCD) scores (7.95 for male and 9.18 for female), indicating better spectral accuracy.

Still, both models faced limitations in generating long or complex utterances due to the lack of extended speech samples in the dataset. FastSpeech 2 also showed some phoneme misalignments, affecting pronunciation stability. These findings highlight the importance of using well-aligned datasets with longer speaker durations and suggest that future work should explore multi-speaker Tacotron 2 models and phoneme-level inputs, while for FastSpeech 2, prosody injection and fine-tuning on a pre-trained English model are recommended to further improve Filipino TTS synthesis.

Acknowledgment

The authors would like to thank Ms. Kiel Gonzales, who provided us with the dictionary used in the MFA pipeline, and to the UP EEEI Digital Signal Processing Laboratory for providing us with the Filipino Speech Corpus. The authors would also like to appreciate the support of their family and friends in the creation of this project.

References

[1] E. J. Renovalles, C. R. Lucas, F. de Leon, A. Aquino and I. Jalandoni, "Text-to-Speech Systems for Filipino Using Unit Selection and Deep Learning," 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Singapore, Singapore, 2021, pp. 212-217, doi: 10.1109/O-COCOSDA202152914.2021.9660431.

[2] T. Gopalakrishnan, S. A. Imam and A. Aggarwal, "Fine Tuning and Comparing Tacotron 2, Deep Voice 3, and FastSpeech 2 TTS Models in a Low Resource Environment," 2022 IEEE International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2022, pp. 1-6, doi: 10.1109/ICDSIS55133.2022.9915932.

[3] M. C. Kelley, S. J. Perry, and B. V. Tucker, "The Mason-Alberta Phonetic Segmenter: a forced alignment system based on deep neural networks and interpolation", Phonetica, vol. 81, no. 5, pp. 451 –508, Sep. 2024, ISSN: 1423-0321. DOI: 10.1515/phon-2024-0015.

[4] C. Tobar, "Text-to-Speech Basics: What Is TTS and Who Uses It?," CourseArc, 2021. https://www.coursearc.com/guest-post-readspeaker-text-to-speech/

[5] C. M. Tatoy, J. L. Pasco, J. I. E. Benedicto and C. R. Lucas, "Harmonic-plus-Noise Network with Linear Prediction and Perceptual Weighting Filters for Filipino Speech Synthesis," 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, 2023, pp. 164–169. doi: 10.1109/SpeD59241.2023.10314926

[6] NVIDIA, "GitHub - NVIDIA/DeepLearningExamples: State-of-the-Art Deep Learning scripts organized by models - easy to train and deploy with reproducible accuracy and performance on enterprise-grade infrastructure.," GitHub, 2018. https://github.com/NVIDIA/DeepLearningExamples/tree/master.

[7] Grapheme-to-Phoneme, "Montreal Forced Aligner 3.0.0 documentation," Montreal Forced Aligner, 2024.

[8] Acoustic Models, "Montreal Forced Aligner 3.0.0 documentation," Montreal Forced Aligner, 2024.

[9] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," Proc. Interspeech, Aug. 2017.

[10] stefantaubert, "GitHub - stefantaubert/mel-cepstral-distance: A Python library for computing the Mel-Cepstral Distance (Mel-Cepstral Distortion, MCD) between two inputs. This implementation is based on the method proposed by Robert F. Kubichek in 'Mel-Cepstral Distance Measure for Objective Speech Quality Assessment'.," GitHub, Apr. 14, 2025. https://github.com/stefantaubert/mel-cepstral-distance.

[11] ming024, "GitHub - ming024/FastSpeech2: An implementation of Microsoft's 'FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,'" GitHub, 2020. https://github.com/ming024/FastSpeech2.

[12] jik876, "GitHub - jik876/hifi-gan: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," GitHub, 2020. https://github.com/jik876/hifi-gan.

[13] ITU-T, "Methods for Subjective Determination of Transmission Quality," ITU-T Recommendation P.800, Aug. 1996. Available: https://www.itu.int/rec/T-REC-P.800-199608-I/en.

[14] The jamovi project (2024). jamovi. (Version 2.6) [Computer Software]. Retrieved from https://www.jamovi.org.