Wav2Vec2-based Automatic Speech Recognition Systems for Filipino Children's Speech

Francesca Jean Santiago
Electrical and Electronics Engineering
University of the Philippines - Diliman
Quezon City, Philippines
francesca.jean.santiago@eee.upd.edu.ph

Glenn L'nore Frilles
Electrical and Electronics Engineering
University of the Philippines - Diliman
Quezon City, Philippines
glenn.lnore.frilles@eee.upd.edu.ph

John Cairu Ramirez
Electrical and Electronics Engineering
University of the Philippines - Diliman
Quezon City, Philippines
john.cairu.ramirez@eee.upd.edu.ph

Rhandley Cajote
Electrical and Electronics Engineering
University of the Philippines - Diliman
Quezon City, Philippines
rhandley.cajote@eee.upd.edu.ph

***Abstract*—Developing automatic speech recognition (ASR) systems for Filipino children's speech remains a significant challenge despite the existing need for automated reading tutors (ARTs) to support early literacy instruction. This paper explores the wav2vec2 adaptation for the Tanglaw corpus, and developed two baseline models: one pre-trained with Filipino adult's speech, and another on English children's speech. It also evaluates two data augmentation strategies which are Spectrogram Augmentation (SpecAugment) and MaskCycleGAN-based voice conversion (MaskCycleGAN-VC) to address the low-resource nature of Filipino children's speech. Across ten model configurations, the best-performing system achieved a Word Error Rate (WER) of 1.80% which substantially outperforms both baseline systems. This result demonstrates an effective approach to developing Filipino children's ASR systems which could particularly be valuable in resource-limited learning environments.**

***Index Terms*—Filipino children ASR, Wav2Vec2, Spectrogram Augmentation, MaskCycleGAN-VC**

I. Introduction

Early reading proficiency remains a critical challenge in the Philippines, where 56.4% of children struggle with age-appropriate literacy, which is a rate significantly higher than regional averages. This crisis stems from systemic barriers including teacher shortages, multilingual complexities, and inadequate resources [1]. While phonological awareness training is essential for reading development [2], conventional teaching methods struggle to provide the personalized feedback required for effective learning.

Recent advances in automatic speech recognition (ASR) show strong potential in enabling automated reading tutors (ART). However, children's speech recognition (CSR) systems face unique obstacles like the acoustic variability of children's speech [3], limited training data for low-resource languages like Filipino, and the disfluent speech patterns children exhibit. Previous Filipino CSR efforts using HMM-GMM frameworks [4], [5] laid important groundwork, though performance was affected by limited training data.

This study introduces two key advancements for Filipino children's speech recognition: (1) utilizing wav2vec2's self-supervised learning framework to address data scarcity [6], and (2) enhancing performance through data augmentation techniques, specifically Spectrogram Augmentation (SpecAugment) and MaskCycleGAN voice conversion (MaskCycleGAN-VC). We established baseline systems using Filipino adult and English children's speech, then fine-tuned using the Tanglaw corpus. The resulting augmentation combinations are systematically evaluated to determine the optimal performance metrics.

The developed system addresses the need to provide scalable literacy support by advancing low-resource ASR techniques, and further establishing the benchmarks for Filipino ART. The integration of this system could reduce dependency on human tutors while offering data-driven insights for classroom instruction which is vital in underserved communities where accessibility of traditional interventions remains an issue.

II. Related Work

Emergent literacy, particularly phonological awareness, is foundational to early reading acquisition among children and is shown to correlate with later reading proficiency [7]. In the Philippines, access to quality early education is hindered by socioeconomic disparities, and many children lack exposure to enriched learning environments [1]. As a response, computer-based instruction has been explored as a more accessible and scalable method of delivering phonological awareness training [4], but the lack of specialized tools for pre-readers in Filipino remains a gap.

ASR converts speech into text and is widely applied in voice assistants, transcription tools, and educational technologies [8]. For CSR, however, the higher variability in pitch, pronunciation, and speaking rate presents challenges that make adult-trained models ineffective [9]. To improve CSR systems, some

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

studies have developed dedicated children's speech corpora, such as the American English PF-STAR and MyST datasets, while local efforts like the Bangsamoro K-3 Assessment Tools (BK3AT) and the Tanglaw corpus have supported Filipino CSR development [10]. Adaptation techniques such as pitch-based augmentation and vocal tract length perturbation (VTLP) have shown improvements in mitigating age mismatch [2], [5].

The architectures of ASR have evolved from traditional HMM-GMM models, which represent phoneme sequences probabilistically, to more sophisticated frameworks like hybrid HMM-DNN and self-supervised models. While HMM-GMM remains widely used, hybrid models like TDNN-HMM offer improved performance through discriminative learning and sequence modeling [11]. A TDNN-HMM model trained on the Filipino Children's Speech Corpus (FCSC) achieved a WER of 0.97%. This is the best result reported, though it is trained and tested on overlapping data [11].

More recently, self-supervised learning methods like wav2vec2 have shown promise in low-resource contexts. Wav2vec2 learns representations from raw audio without requiring extensive labeled data and can be fine-tuned with domain-specific corpora [12]. This method has been effective even with limited Filipino speech data [13] and outperformed HMM-based systems in several English CSR studies [14]. The multilingual XLS-R variant, pre-trained on over 400k hours of speech in 128 languages, further enhances performance by capturing cross-linguistic phonemic features [15].

Data augmentation plays a key role in improving CSR systems, especially when working with limited datasets. SpecAugment applies transformations like time warping and frequency masking to diversify training data and has been shown to reduce overfitting and improve WER in noisy or variable conditions [16]. MaskCycleGAN-VC, a GAN-based conversion technique, enables speaker adaptation without parallel data and has demonstrated better performance than earlier CycleGAN-based models [17].

Results across various Filipino ASR and CSR systems show that wav2vec2-based models tend to outperform HMM-based approaches in terms of WER when combined with proper data augmentation methods [5], [14]. While HMM-based models remain common in Filipino CSR, the increasing use of wav2vec2 in other languages points to a shift toward more robust architectures.

III. Methodology

This study investigated the use of Filipino children's speech and data augmentation techniques for fine-tuning two baseline systems. The first baseline, derived from Filipino adult speech (FA), was expected to align well with Filipino phonetic patterns but introduced an age mismatch with the target child speech. In contrast, the second baseline, based on English children's speech (EC), mitigated the age mismatch but lacked familiarity with Filipino phonology. Both systems were employed to compare their effectiveness in developing a Filipino CSR system.

TABLE I
Overall Dataset Description

Usage	Dataset	Duration	Speaker Information
Baseline	FSC (from LDC)	5.06 hours	Filipino Adult
Augmentation		0.17 hours	
Baseline	MyST (from LDC)	84.2 hours	English Children
Augmentation		0.17 hours	
Augmentation	Tanglaw (from UP DSP)	0.17 hours	Filipino Children
Fine-tuning		42.81 hours	
Testing		4.76 hours	

A. Data and System Acquisition

The corpora used in this study were sourced from the University of the Philippines Digital Signal Processing Laboratory (UP DSP) and the University of Pennsylvania Linguistic Data Consortium (LDC). A summary of these datasets is provided in Table I. The train data for the FA baseline were sourced from the Filipino Speech Corpus (FSC), which contained utterances from speakers aged 20 to 60. Of this train set, 2.49 hours were female speech and 2.57 were male speech. The train data for the EC baseline were sourced from the My Science Tutor Corpus (MyST), which consisted of conversational speech from students aged 8 to 10. For fine-tuning both baselines, the Tanglaw Corpus was acquired. This comprised speech from early-reading Filipino children, including 28.44 hours of female speech and 19.13 hours of male speech. For data augmentation, 0.17 hours were sampled from each of the three corpora. As of writing, the Tanglaw Corpus has not been made available to the public.

A portion of the Tanglaw Corpus was reserved for the test set, which comprise 4.76 hours of previously unseen speech data. This set featured recordings from early Filipino readers and included both clean and disfluent utterances. Test cases with repetitions, hesitations and reading miscues, as well as background noise accounted for 14.45% of the data. This composition allowed for a more realistic evaluation of ASR performance under conditions reflective of actual early reading behavior.

Existing pre-trained models were acquired to develop the CSR systems. The EC baseline was adapted from [18], wherein 960 hours of unlabeled English speech were used to pre-train the model before fine-tuning on MyST. Conversely, the XLS-R-300M[1] model, pre-trained on multiple languages including Filipino, was acquired to develop the FA baseline.

B. Data Preprocessing

To fit the wav2vec2 architecture, all speech data were split into 10-20 second clips and sampled at 16 KHz. The corresponding transcripts were converted into lowercase and with no punctuations.

[1] https://huggingface.co/facebook/wav2vec2-xls-r-300m

TABLE II
PARAMETERS FOR SPECAUGMENT

`apply_spec_augment`	True
`mask_time_prob`	0.05
`mask_time_length`	10
`mask_time_min_masks`	2
`mask_feature_prob`	0.0

TABLE III
HYPERPARAMETERS FOR THE MASKCYCLEGAN MODEL

Batch size	1
Generator learning rate	5e-4
Discriminator learning rate	5e-4
Learning rate decay	1e4
Weight (λ_{cyc})	10
Weight (λ_{id})	5

C. Data Augmentation

1) Spectrogram Augmentation: Traditional implementations of SpecAugment operate on log-mel spectrograms, where time and frequency masking are applied to augment the input features [16]. In contrast, wav2vec2 processes raw audio waveforms directly. To incorporate SpecAugment in this context, the Hugging Face implementation of wav2vec2 provides the `apply_spec_augment` parameter, which applies SpecAugment internally during training[2]. Among the available SpecAugment settings, only a limited subset is applied in this study, as shown in Table II. With these settings, only time masking is applied, as the frequency masking implementation operates on CNN feature maps rather than true frequency bands, which may hurt training performance on smaller datasets such as Tanglaw.

The time masking is implemented by applying fixed-length (`mask_time_length`) time masks to the feature sequence, wherein the probability of starting a time mask at each time step is defined by `mask_time_prob`. To ensure sufficient augmentation on shorter inputs, `mask_time_min_masks` guaranteed a minimum of 2 time masks per sample.

2) MaskCycleGAN Voice Conversion: MaskCycleGAN-VC, a state-of-the-art method for nonparallel voice conversion using CycleGAN [3], was implemented using Python on a Windows-based system. The model is generally trained on two distinct datasets and generates audio waveforms that retain the linguistic content of one (source) while adopting the speech characteristics of the other (target). Its training incorporates a temporal masking strategy applied to the input mel-spectrogram to enhance the model's ability to reconstruct masked segments. The architecture comprises two neural networks: a generator, which produces generated data, and a discriminator, which attempts to distinguish between real and generated data. Through adversarial training, the generator is refined based on feedback from the discriminator until the synthetic outputs become indistinguishable from real data. In this study, MaskCycleGAN-VC is implemented for two source-target pairs: FSC-Tanglaw and MyST-Tanglaw. The hyperparameters used for training the model are outlined on Table III as adapted from the original repository[3].

D. Training

Training was conducted in two phases. In the first phase, the EC and FA baseline models were trained. In the second phase, data augmentation techniques were incorporated in fine-tuning these baselines using the Tanglaw corpus. For comparison, the baselines were also fine-tuned on the Tanglaw corpus without any data augmentation. This process resulted in a total of eight fine-tuned models for testing and evaluation, as illustrated in Figure 1. All CSR systems were implemented using the wav2vec2 model of Hugging Face, trained on Google Colab using an A100 GPU. The hyperparameters used for training followed the recommendations of the official documentation[4] and were adjusted according to validation loss and evaluation word error rate per model. To prevent overfit, early stoppage was implemented to halt training once the evaluation results stagnate.

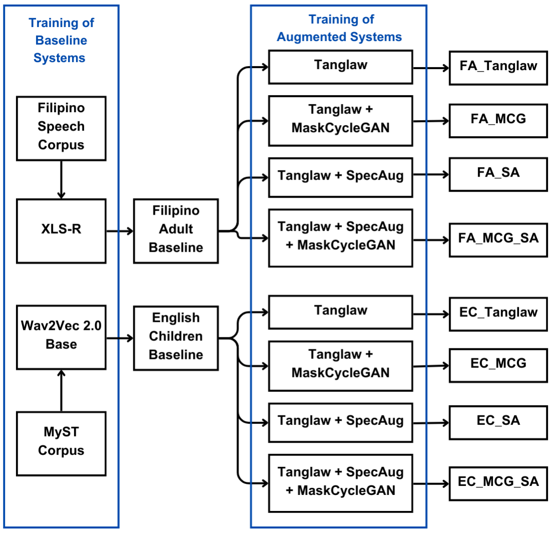

Fig. 1. Training block of the baseline and augmented systems.

E. Testing

Word Error Rate (%WER) is the standard metric used to evaluate the performance of ASR systems. It is the ratio of the number of errors in the predicted text to the total number of words (N) in the actual text, effectively measuring the accuracy of transcriptions generated by the ASR system. The defined errors are: insertions (I), words added in the transcription but not spoken; substitutions (S), spoken words transcribed incorrectly as different words; and deletions (D), spoken words omitted from the transcription. The formula for %WER is

[2] https://huggingface.co/docs/transformers/en/model_doc/wav2vec2
[3] https://github.com/GANtastic3/MaskCycleGAN-VC
[4] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

shown in (1), which was implemented in this study through Hugging Face's built-in evaluate library[5].

$$\%WER = \frac{I+S+D}{N} \times 100\% \quad (1)$$

The same test set was used to obtain the corresponding %WER for all the developed systems. The performance of the baseline and augmented systems were then compared using their resulting %WER, where lower %WER indicated better transcription accuracy. The system with the lowest WER was considered the most effective in transcribing Filipino children's speech. Furthermore, adapted from [5], the relative improvement (%RI) of each augmented system (S_A) compared to its baseline system (S_B) was computed. The formula for %RI is shown in (2).

$$\%RI = \frac{|S_A - S_B|}{S_B} \times 100\% \quad (2)$$

IV. Results and Analysis

A. Main Results

The main results of the study are summarized in Table IV.

1) Baseline Models: The XLSR-300M pretrained model was fine-tuned on FSC for 15 epochs to develop the FA baseline. Upon testing, the FA baseline achieved a WER of 39.70%. Meanwhile, the acquired EC baseline[6] achieved a WER of 100.4%. Since the EC baseline was pretrained on English adult speech and fine-tuned on the MyST corpus, it lacked exposure to Filipino speech. As a result, it produced a high WER, which exceeded 100% due to the number of insertions, deletions, and substitutions in the predicted transcription surpassing the total number of words in the reference text.

To assess the impact of data augmentation, the FA and EC baselines were both fine-tuned to the Tanglaw corpus and were recorded as FA_Tanglaw and EC_Tanglaw respectively. FA_Tanglaw was trained for 9 epochs and achieved a WER of 2.6%, 93.45% RI over the FA baseline. EC_Tanglaw was trained for 6 epochs and achieved a WER of 2.5%, 97.51% RI over the EC baseline.

2) MaskCycleGAN-Voice Conversion: The portioned subsets of FSC, MyST, and Tanglaw for data augmentation were converted into mel spectrograms using a MelGAN vocoder[7], which enabled the conversion of synthesized speech back to WAV format after inference. Two voice conversion models were trained: one using the FSC-Tanglaw pair and the other using the MyST-Tanglaw pair. Both models were trained for 6172 epochs with results shown on Table V.

For the MyST-Tanglaw model, the discriminator loss began to approach zero after the 3100th epoch, then spiked again beyond the 3500th epoch. This pattern indicated that the discriminator had started to overpower the generator, disrupting the adversarial balance crucial for stable training. As a result, training became unstable beyond this point. Therefore, the checkpoint at the 3100th epoch was selected as the best-performing model, as it represented the last stable state before the onset of instability. This is justified by [5], which stated that training is considered stable and suitable for evaluation once the discriminator loss consistently falls below 0.05.

Both models were used to generate synthesized speech targeting the speech characteristics of the Tanglaw corpus which produced 0.17 hours of MyST-converted speech and 0.56 hours of FSC-converted speech. These augmented sets were then combined with the original Tanglaw corpus to train the EC_MCG and FA_MCG systems, respectively. The EC_MCG model, trained for 19 epochs, achieved a WER of 1.90%, representing a 98.11% RI over the EC baseline. Similarly, the FA_MCG model, trained for 7 epochs, achieved a WER of 3.90%, corresponding to a 90.17% RI over the FA baseline.

3) Spectrogram Augmentation: Both the FA and EC baselines were fine-tuned on the Tanglaw corpus with SpecAugment applied during training. The FA baseline was trained for 15 epochs to produce the FA_SA model, which achieved a WER of 1.80% with a 95.47% RI. Similarly, the EC baseline was trained for 6 epochs to produce the EC_SA model, which achieved a WER of 3.80% with a 96.22% RI.

4) Combination of Augmented Data: In combining the two data augmentation techniques, the FA baseline was fine-tuned on FSC-Tanglaw synthesized speech and the Tanglaw corpus for 12 epochs with SpecAugment applied. This produced the FA_MCG_SA model. The EC baseline was fine-tuned for 19 epochs in the same manner using the MyST-Tanglaw synthesized speech and Tanglaw corpus, which produced the EC_MCG_SA model. The FA_MCG_SA model achieved a WER of 2.00% with a 94.96% RI from the FA baseline, while the EC_MCG_SA model achieved a WER of 2.67% with a 97.34% RI from the EC baseline.

B. Discussion of Results

1) Filipino Adult Baseline: The poorest performance of the FA baseline on Filipino children's speech was a WER of 39.70%, demonstrating the significant mismatch between adult and child speech characteristics. After applying SpecAugment, the FA_SA model yielded a WER of 1.80% which is the best performance among all systems. This result suggests that while the FA baseline benefits from linguistic familiarity with Filipino phonemes, it is sensitive to child-specific acoustic variability without augmentation. SpecAugment appears especially effective in addressing this mismatch likely due to its ability to simulate realistic variations in pitch and timing of young speakers. Although combining both augmentation methods in the FA_MCG_SA model still led to strong results, the marginal difference compared to FA_SA suggests less performance gains when the base model is already well-aligned phonetically.

2) English Children Baseline: The EC baseline initially yielded the highest WER at 100.40%, primarily due to its lack

[5] https://huggingface.co/docs/evaluate/en/index

[6] https://huggingface.co/lijialudew/wav2vec_children_ASR

[7] https://github.com/descriptinc/melgan-neurips

TABLE IV
SUMMARY OF RESULTS

Type	Model ID	Pretrained Model	Fine-tuning Dataset	Data Augmentation	Test WER%	RI%
Baseline	FA	XLSR-300M	FSC	N/A	39.70%	N/A
Fine-tuned	FA_Tanglaw	FA	Tanglaw	N/A	2.60%	93.45%
Fine-tuned	FA_MCG	FA	Tanglaw, FSC	MaskCycleGAN-VC	3.90%	90.17%
Fine-tuned	FA_SA	FA	Tanglaw	SpecAugment	1.80%	95.47%
Fine-tuned	FA_MCG_SA	FA	Tanglaw, FSC	MaskCycleGAN-VC, SpecAugment	2.00%	94.96%
Baseline	EC	Wav2Vec2 Base	MyST	N/A	100.40%	N/A
Fine-tuned	EC_Tanglaw	EC	Tanglaw	N/A	2.50%	97.51%
Fine-tuned	EC_MCG	EC	Tanglaw, MyST	MaskCycleGAN-VC	1.90%	98.11%
Fine-tuned	EC_SA	EC	Tanglaw	SpecAugment	3.80%	96.22%
Fine-tuned	EC_MCG_SA	EC	Tanglaw, MyST	MaskCycleGAN-VC, SpecAugment	2.67%	97.34%

* The best-performing models for the FA and EC baselines, as determined by WER%, are highlighted in yellow.

TABLE V
SUMMARY OF TRAINING FOR MASKCYCLEGAN

Conversion Model	Checkpoint	Discriminator Loss	Generator Loss
FSC-Tanglaw	6120	0.0040	6.3429
MyST-Tanglaw	3100	0.0027	7.0235

of exposure to Filipino phonemes. However, after fine-tuning on the Tanglaw corpus augmented with voice-converted MyST data, the performance of the EC_MCG model drastically improved to a WER of 1.90%. This result demonstrates that age-aligned acoustic features, even when drawn from a different language, can be used effectively through voice conversion to reduce domain mismatch. In addition to the augmentation, the introduction of a custom tokenizer built from the Tanglaw corpus likely contributed to the improved performance by better aligning the model's output vocabulary with Filipino orthographic patterns. While EC_SA also showed improvement, it was outperformed by EC_MCG, indicating that voice conversion performs better in compensating for the initial phonetic mismatch.

3) Sensitivity to Disfluencies: For reading support applications, it is crucial for ASR systems to detect disfluencies like miscues, stuttering, and hesitations to aid learning progress. Table VI presents representative test set cases with predictions from each augmented system. Baseline systems were excluded due to unintelligible outputs. As shown, all augmented systems were successful in identifying the intended speech despite background noise and/or speaker hesitation. In cases involving stuttering, the FA-based systems, FA_SA, FA_MCG, and FA_MCG_SA, demonstrated a higher sensitivity to repeated phonemes, while, only EC_MCG_SA was able to identify stuttering among the EC-based systems, suggesting the importance of multilingual pretraining and data augmentation in improving disfluency detection. Notably, only the FA_SA model was able to accurately transcribe repeated syllables such as "ahahahahaha," demonstrating its robust handling of complex speech patterns.

C. This study in the context of previous Filipino ASR

This work highlights a shift in Filipino ASR development from conventional statistical models toward self-supervised, data-efficient architectures. While earlier systems often relied on HMM-based methods, their limitations became more pronounced in low-resource and acoustically mismatched scenarios, particularly for children's speech. The consistent improvement observed across all wav2vec2-based models in this study, especially when augmented with spectrogram-based and generative techniques, demonstrates the potential of modern ASR approaches to overcome long-standing challenges in resource-limited languages.

In doing so, this work also underscores the value of strategic model pretraining and augmentation over large data volumes. Rather than requiring extensive annotated corpora, the combination of self-supervised learning and targeted data augmentation proved effective in adapting to highly variable speech patterns such as those found in early Filipino readers.

V. CONCLUSION AND RECOMMENDATIONS

A. Conclusion

Based on the training and evaluation of ten wav2vec2-based ASR models, the study examined different strategies to improve speech recognition for Filipino children. Two baseline systems were developed using pre-trained English children's speech and Filipino adult speech, respectively. These baselines were then fine-tuned with two different data augmentation techniques, Spectrogram Augmentation (SpecAugment) and MaskCycleGAN-based voice conversion (MaskCycleGAN-VC), both individually and in combination. Among all configurations, the best-performing model was the Filipino adult baseline fine-tuned with SpecAugment, achieving a word error rate (WER) of 1.80% on the test set of early readers in the Filipino children's speech corpus, Tanglaw. This result marks a substantial improvement from the original Filipino adult baseline WER of 39.70% and the English children baseline WER of 100.40%. Notably, all augmented models achieved better performance than their respective baselines, demonstrating the effectiveness of both augmentation techniques in addressing the data limitations of Filipino children's speech. The results in Table IV reflect consistent performance gains with augmentation, particularly when using SpecAugment for the Filipino adult baseline, and MaskCycleGAN-VC for the English children baseline. This emphasizes the importance of

TABLE VI
DISFLUENCY OUTPUTS AND OBSERVATIONS

Reference Text	Disfluency	EC_MCG_SA	EC_MCG	EC_SA	EC_Tanglaw	FA_MCG_SA	FA_MCG	FA_SA	FA_Tanglaw
naiinis na ako sa suysoy mo	notable background noise	matched reference	matched reference	matched reference	matched reference	matched reference	matched reference	matched reference	matched reference
..labada ahahahahahaha	..labada ahahahahaha	..labada ahaha	..labada hahaha	..labada ahahahaha	..labada ahaha	..labada ahahahaha	..labada ahahahaha	..labada ahahahahaha	..labada ahahahaha
..hanggang sa mawalan ng malay	maw-mawalan ng malay	matched reference	..sama mawalan ng..	matched reference	matched reference	..sa ma mawalan..	..sa ma mawalan..	..sa maw mawalan..	matched reference
ang mabait na kalabaw ay mapagbigay	mapag.. bigay	matched reference	matched reference	matched reference	matched reference	matched reference	matched reference	matched reference	matched reference

selecting models and data augmentations tailored to the target domain, especially in low-resource, high-variability settings.

B. Recommendations for Future Work

This work suggests several research extensions for future work. First, we recommend exploring using multilingual wav2vec2 models with broader Filipino language coverage, including regional dialects. Further investigation into other generative voice conversion methods may also yield improved data augmentation quality. In addition, evaluating the system on spontaneous or conversational Filipino children's speech can provide insight into its performance beyond scripted reading tasks. Practical implementation can also focus on embedding the optimized ASR within an interactive learning platform to deliver real-time phonological feedback. This integration has the potential to significantly enhance literacy instruction by providing adaptive, technology-mediated tutoring.

REFERENCES

[1] V. G. Ulep, K. L. Dela Luna, J. Bagas, J. P. Mendoza, A. C. Manuel, and L. D. Casa, "Behind the slow start: An assessment of early childhood care and development in the philippines," 2024.

[2] J. R. Maranan, "An automated speech recognition system for phonological awareness of kindergarten students in filipino," in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2022, pp. 1–7.

[3] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100–2104.

[4] F. P. Tupas and M. Linas-Laguda, "Blended learning – an approach in philippine basic education curriculum in new normal: A review of current literature," *Universal Journal of Educational Research*, vol. 8, pp. 5505–5512, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229010157

[5] C. Abion, N. C. Lumapag, J. C. Ramirez, C. Resulto, and C. R. Lucas, "Comparison of data augmentation techniques on filipino asr for children's speech," in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2023, pp. 60–65.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[7] C. Lonigan, S. Burgess, and J. Anthony, "Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study," *Developmental psychology*, vol. 36, pp. 596–613, 09 2000.

[8] M. Helander, T. S. Moody, and M. G. Joost, "Chapter 14 - systems design for automated speech recognition," in *Handbook of Human-Computer Interaction*, M. HELANDER, Ed. Amsterdam: North-Holland, 1988, pp. 301–319. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444705365500191

[9] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of asr technologies for children's speech," *Proceedings of the 2nd Workshop on Child, Computer and Interaction, WOCCI '09*, 11 2009.

[10] K. D. Gonzales, J. R. Maranan, F. P. D. Santelices, E. J. M. Renovalles, N. D. Macale, N. A. A. Palafox, and J. M. A. Mendoza, "BK3AT: Bangsamoro k-3 children's speech corpus for developing assessment tools in the bangsamoro languages," in *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, M. Melero, S. Sakti, and C. Soria, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 59–65. [Online]. Available: https://aclanthology.org/2024.sigul-1.8

[11] J. A. Y. Ing, R. M. Pascual, and F. D. Dimzon, "A hybrid tdnn-hmm automatic speech recognizer for filipino children's speech," in *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 2022, pp. 1–6.

[12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[13] J. Goma, J. Alberto, K. Antonio, and P. Pedro, "Speech recognition of tagalog talisay batangueño accent in the philippines using wav2vec2.0," in *IC4E '24'*, 09 2024, pp. 416–421.

[14] R. Jain, A. Barcovschi, M. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition." *IEEE Access*, vol. PP, pp. 1–1, 01 2023.

[15] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," 2021. [Online]. Available: https://arxiv.org/abs/2111.09296

[16] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, ser. interspeech_2019. ISCA, Sep. 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[17] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, and M. Wieling, "Making more of little data: Improving low-resource automatic speech recognition using data augmentation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 715–729. [Online]. Available: https://aclanthology.org/2023.acl-long.42

[18] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Analysis of self-supervised speech models on children's speech and infant vocalizations," in *IEEE Workshop on Self-Supervision in Audio, Speech and Beyond (SASB)*, 2024.