IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.19, No.2, April 2025, pp. 165~176
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.104646

The Impact of Data Augmentation Techniques on Improving Speech Recognition Performance for English in Indonesian Children Based on Wav2Vec 2.0

Maimunah Maskur*[1], Amalia Zahra[2]
[1,2]Bina Nusantara University, Jakarta, Indonesia
e-mail: ***[1]maimunah001@binus.ac.id**, [2]amalia.zahra@binus.edu

Abstrak

Pendidikan usia dini adalah fase krusial dalam pembentukan karakter dan kemampuan bahasa anak. Penelitian ini mengembangkan model Automatic Speech Recognition (ASR) untuk mengenali ucapan anak-anak Indonesia dalam Bahasa Inggris. Proses dimulai dengan pengumpulan dan pemrosesan dataset rekaman suara anak-anak, yang kemudian diperluas melalui teknik augmentasi data guna meningkatkan variasi pengucapan. Model ASR Wav2Vec 2.0 yang telah dilatih sebelumnya difine-tune dengan dataset asli dan augmentasi. Evaluasi menggunakan Word Error Rate (WER) dan Character Error Rate (CER) menunjukkan peningkatan akurasi signifikan, dengan WER turun dari 53% menjadi 45% dan CER dari 33% menjadi 27%, mencerminkan peningkatan kinerja sekitar 15%. Analisis lebih lanjut mengungkap kesalahan pelafalan pada fonem seperti /ð/, /θ/, /r/, /v/, /z/, dan /ʃ/, yang tidak umum dalam Bahasa Indonesia, muncul dalam bentuk substitusi, penghilangan, atau penambahan fonem pada kata-kata seperti "three," "that," "rabbit," "very," dan "zebra." Temuan ini menyoroti perlunya latihan fonem spesifik, pendekatan berbasis audio dengan umpan balik ASR, serta teknik listen and repeat dalam pengajaran Bahasa Inggris kepada anak-anak.

Kata kunci*—Pendidikan usia dini, Automatic Speech Recognition, Augmentasi, Character Error Rate, Word Error Rate*

Abstract

Early childhood education is crucial in shaping children's character and language skills. This study develops an Automatic Speech Recognition (ASR) model to recognize the speech of Indonesian children speaking English. The process begins with collecting and processing a dataset of children's speech recordings, which is then expanded using data augmentation techniques to enhance pronunciation variations. The pre-trained ASR Wav2Vec 2.0 model is fine-tuned with the original and augmented datasets. Evaluation using Word Error Rate (WER) and Character Error Rate (CER) shows a significant accuracy improvement, with WER decreasing from 53% to 45% and CER from 33% to 27%, reflecting a performance increase of approximately 15%. Further analysis reveals pronunciation errors in phonemes such as /ð/, /θ/, /r/, /v/, /z/, and /ʃ/, which are uncommon in the Indonesian language, manifesting as substitutions, omissions, or additions in words like "three," "that," "rabbit," "very," and "zebra." These findings highlight the need for targeted phoneme training, audio-based approaches with ASR feedback, and the listen-and-repeat technique in English language instruction for children.

Keywords— *Early childhood education, Automatic Speech Recognition, Augmentation, Character Error Rate, Word Error Rate*

Received February 11th,2025; Revised Aprtil 26th, 2025; Accepted April 29th, 2025

1. INTRODUCTION

Early childhood education is fundamental in shaping children's character and cognitive abilities. According to PUSDATIN KEMDIKBUD (2022), children's sensitivity to language is exceptionally high during the golden period of brain development (ages 0-5). At this stage, they rapidly absorb and learn language skills, including number recognition, pronunciation, and socialization. Therefore, early childhood education (PAUD) plays a crucial role and requires special attention to optimize its impact on character development.

Technological advancements have made technology-based learning methods, such as multimedia and computer applications, integral to education. These innovations enhance the quality and accessibility of learning (Jasnanto, 2022). One commonly used multimedia format in teaching is educational games, especially for young children. Sandri et al. (2019) developed an educational game for preschoolers to learn basic English. Their study found that preschoolers recognized objects such as fruits, colors, shapes, and letters faster when using the game. This was due to the application's audio feature, which allowed children to hear and repeat English pronunciations when encountering difficulties.

Similarly, Tahapary and Wahyono (2022) developed a mobile-based English learning application for elementary school students. This application helps children understand numbers, letters, daily sentences, family members, and objects in English, improving their pronunciation and comprehension.

Children often use imperfect language structures with limited vocabulary and phoneme articulation in communication. This is influenced by age-related factors, as their speech-motor system—including vocal muscles, teeth structure, jaw shape, and tongue flexibility—has not fully developed (Fitriana, 2021).

To effectively implement technology-based English learning, a system capable of evaluating children's pronunciation is necessary. In this context, speech recognition plays a vital role, and Automatic Speech Recognition (ASR) offers a relevant solution. ASR is a technology that converts human speech into text, potentially enhancing English learning for children.

Chen et al. (2023) at Oregon Health & Science University, Portland, USA, have researched ASR. They optimized the Wav2Vec 2.0 model for ASR with limited training data, specifically for preschool children with speech impairments. Their study analyzed attention patterns in the pre-trained Wav2Vec 2.0 model using two optimization techniques: local attention mechanisms and parameter sharing across blocks. These techniques improved absolute WER by 1.8% on the dev-clean dataset and 1.4% on the test-clean dataset.

Several studies have explored pre-trained ASR models like Wav2Vec base and XLSR-53 for adult speech recognition in Indonesia. Syahputra and Zahra (2021) developed an Indonesian ASR model using a Wav2Vec base, achieving a WER of 21%. Meanwhile, Arisaputra and Zahra (2022) used XLSR-53, which supports multiple languages, including Indonesian, achieving an optimal WER of 7% across three datasets (TITML-IDN, Magic Data, and Common Voice).

In education, ASR has significant potential to assist children in learning English by providing real-time feedback on their pronunciation. The rapid development of ASR technology, driven by deep learning, has led to adopting pre-trained models such as Wav2Vec 2.0 (Baevski et al., 2020). This model, trained on large datasets, captures robust and general acoustic features, making it adaptable for specific ASR tasks with limited data through fine-tuning.

This study utilizes the Wav2Vec 2.0 model to build an ASR system for Indonesian children's English speech recognition. However, a major challenge is the limited availability of children's speech datasets. Data augmentation techniques are applied to enhance model robustness, balance data distribution, and reduce overfitting, as demonstrated in previous research (Park et al., 2019; Chen et al., 2020). Thus, this study implements data augmentation to

improve the fine-tuning performance of the Wav2Vec 2.0 model in recognizing English speech from Indonesian children. Additionally, model evaluation is conducted using Word Error Rate (WER) and Character Error Rate (CER), two metrics that provide deeper insights into the model's speech recognition performance**.**

The results of this study are expected to contribute to the development of more accurate ASR models, particularly for children's speech. Additionally, it aims to provide an effective learning strategy to help improve the pronunciation skills of Indonesian children in English sentence articulation, specifically through the identification of pronunciation errors at the phoneme and word levels.

2. METHODS

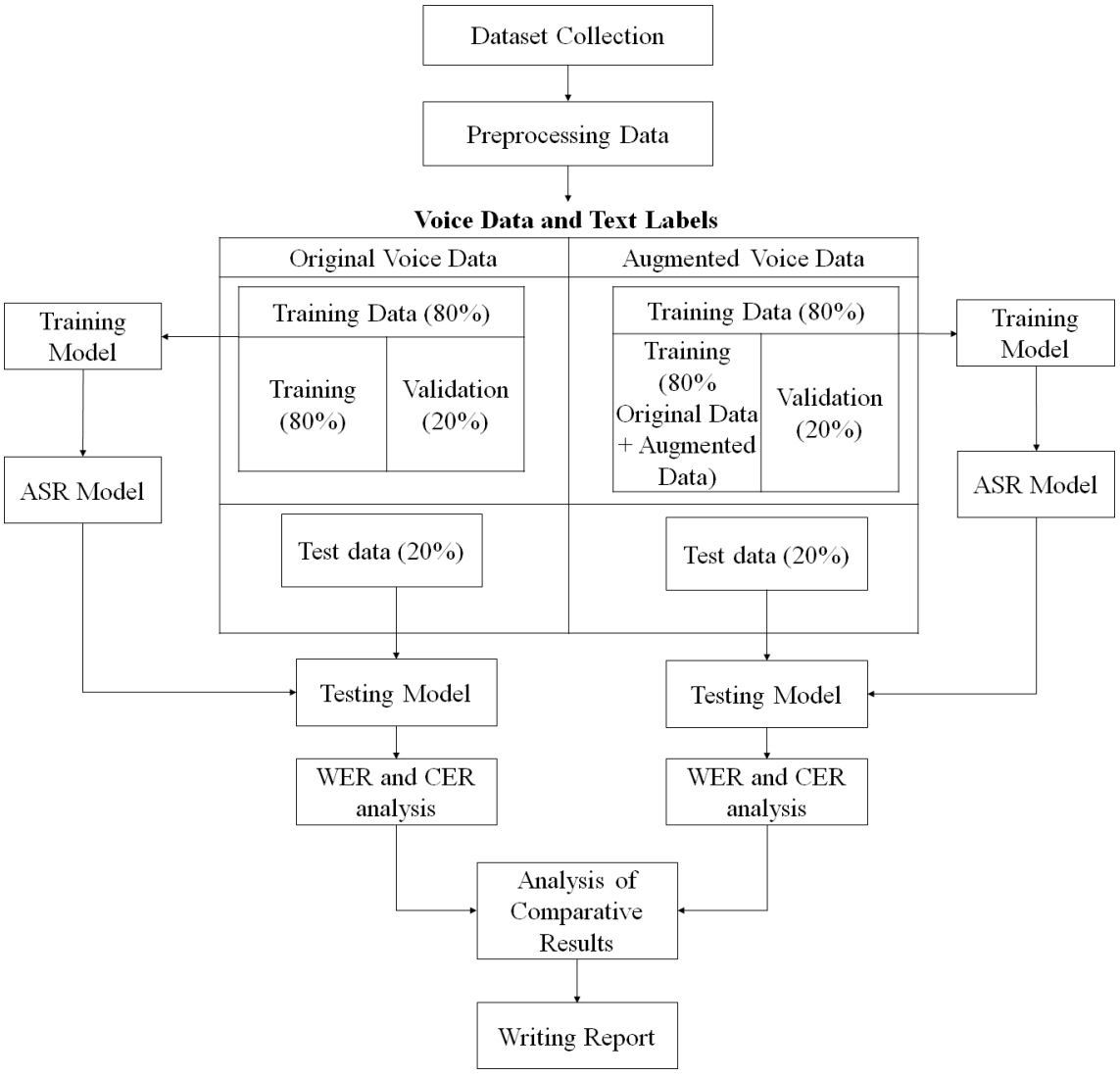

Figure 1 Research Stages

In Figure 1, the research stages illustrate the entire research process that has been conducted and consists of the following elements:

The research began with collecting voice recordings from Indonesian children, specifically those aged 4-8 years who speak English. These children were primary school (Grade 1) and kindergarten students.

Once the data was collected, a preprocessing step prepared it for model training. This involved audio normalization and transcription labeling. The data was split into 80% for training and 20% for testing. The training data was split into two subsets, with 80% used for model training/fine-tuning and 20% used for model validation.

The next step involved fine-tuning the Wav2Vec 2.0 model. Fine-tuning was conducted twice, using two different datasets: the original recordings of Indonesian children and the augmented dataset. This approach aimed to assess the impact of augmentation techniques on improving the model's ability to recognize children's speech, especially with limited data.

After fine-tuning, the model was evaluated using Word Error Rate (WER) and Character Error Rate (CER). WER assessed word-level errors in the model's transcription, while CER measured errors at the character level. These evaluations were conducted on a separate test set to gauge the model's accuracy on entirely new data.

The ASR model trained with both data types and techniques was then tested to generate transcription predictions. WER and CER were recalculated at this stage to evaluate the model's performance.

Finally, the WER and CER results from both models were analyzed and compared to assess the effectiveness of each technique in improving ASR model performance. This comparative analysis aimed to understand how data augmentation can positively impact the accuracy of children's speech recognition in a model trained with limited data. Further details regarding the research process can be found in the subsection below.

2.1 Data Collection and Preprocessing

This study's data collection involved gathering voice recordings of Indonesian children speaking English. The dataset was created from audio recordings of children aged 4-8 years, sourced from two kindergartens and one primary school. A total of 70 children recited 192 short English sentences, with an average of 1.6 repetitions per sentence.

The children varied in age, with thirty-eight boys (54.3%) and thirty-two girls (45.7%). The age distribution included twenty-two children aged 7 (31.4%), ten aged 8 (14.3%), three aged 4 (4.3%), six aged 5 (8.6%), twenty-seven aged 6 (38.6%), and only two children aged 7 (2.9%). Altogether, 316 audio files were produced, totaling 9 minutes and 42 seconds.

For initial data preprocessing, all recordings were converted to the .wav format. This format was chosen because WAV is uncompressed, preserving the original audio quality. Additionally, the recordings were converted to consistent technical specifications: Mono channels, 16,000 Hz sample rate, and Signed 16-bit PCM encoding. This standardization ensured consistent audio quality, which is essential for accurate ASR model training.

After format conversion, manual transcription labeling was completed. This transcription was the reference text for training and testing the ASR model.

2. 2 Data Augmentation

In the study conducted by Chen et al. (2021), data augmentation was used to improve speech recognition for children, particularly due to the limited training data and the lack of children's speech data. Data augmentation helps enhance the model's robustness to the acoustic variability of children's speech. The augmentation techniques tested included pitch perturbation, speech perturbation, tempo perturbation, volume perturbation, reverberation augmentation, and spectral augmentation. This study performed volume perturbation by altering the audio volume using factors ranging from 0.125 to 2. In contrast, pitch perturbation was applied to adult speech to make it closer to children's speech. The results showed that data augmentation improved children's speech recognition performance, with a Character Error Rate (CER) of 18.82% and

16.53% on two evaluation sets.

In this study, several augmentation techniques were applied to the audio data to enrich the variety in the Indonesian children's speech dataset. The augmentation process involved two types of modification techniques: pitch perturbation and volume perturbation.

Pitch perturbation involves shifting the pitch of the audio by ±200 cents (equivalent to ±2 semitones), providing variation in pitch without altering the duration. The `shift_pitch` function adjusts the audio frequency by resetting the frame rate according to the specified cent value.

Volume perturbation changes the volume of the audio in two levels: a decrease of -4 dB and an increase of +4 dB. This technique aims to simulate volume variations that may occur in real-world environments, which the ASR model might encounter. Increasing or decreasing the dB of the original audio volume generates variations that help the model recognize speech at different intensity levels.

The augmented data consists of 201 audio files from the training set. Each audio file was given two new versions with altered pitch and volume and saved with names indicating the applied shift values.

The naming convention for the augmented files is as follows: audio-01_pitch_-200.wav, audio-01_pitch_200.wav, audio-01_volume_-4.wav, and audio-01_volume_4.wav. The audio augmentation process resulted in 1005 new files, which were then combined with the original files. The merged files will be used as training data for the ASR model.

The augmentation process generated 1,005 new audio files, which were then combined with the original files to form an expanded training dataset. This study ensured dataset consistency by normalizing text labels to maintain uniform labels across the entire training dataset, including original and augmented data. Consistent tokenization was also applied, and periodic WER and CER evaluations were conducted on the validation set. The results showed that using a mixed dataset—combining original data with augmented data from volume perturbation and pitch perturbation—enabled the Wav2Vec 2.0 model training process to run effectively, as indicated by valid and consistent values of Training Loss, Validation Loss, WER, and CER throughout each training step and epoch.

2. 3 Wav2Vev 2.0

Wav2Vec 2.0 is a model Facebook AI Research (FAIR) developed to extend the previously introduced Wav2Vec model. Wav2Vec 2.0 is a framework for self-supervised learning of speech representations. This framework uses a masking technique on raw waveform representations and performs contrastive tasks on quantized speech representations. In their research, Baevski et al. (2019) explained that pre-training on unlabeled data has significant potential for speech processing. This approach is effective when a large amount of labeled data is available. The framework is expected to have a broader impact on speech recognition worldwide, particularly for languages that lack speech recognition technology, as current systems require hundreds or thousands of hours of labeled data, which are difficult to collect for most languages.

In their experiment, Baevski et al. (2019) used only 10 minutes of labeled training data or 48 recordings with an average duration of 12.5 seconds. The framework achieved a word error rate (WER) of 4.8/8.2 on the test-clean/other sets of Librispeech. This result set a new state-of-the-art benchmark for noisy speech recognition on Librispeech. In the 100-hour clean Librispeech setup, Wav2Vec 2.0 outperformed the previous best results using 100 times less labeled data.

Baevski et al. (2019) also noted that this framework could be used to improve the quality of speech transcription, reduce annotation costs, and enhance the performance of speech recognition systems. It can also be applied to improve speech recognition systems for languages that do not have sufficient speech recognition technology.

2. 4 Fine Tuning Model Wav2Vec 2.0

The evaluation methods used in this research is based on Word Error Rate (WER) and Character Error Rate (CER). The experiments conducted in this study involved fine-tuning the Wav2Vec 2.0 model using two distinct datasets. The first dataset comprised 316 samples, divided into 201 training samples, 51 validation samples, and 64 testing samples, sourced from Indonesian children's speech recordings. The second dataset was an augmented version comprising 1120 samples, including 1005 training samples (a combination of original and augmented data), 51 validation samples, and 64 testing samples.

During fine-tuning, the same set of hyperparameters was applied to both datasets, although the training configurations, such as the number of *epochs*, were adapted for each experiment, with *epoch*s 100, 200, and 300.

```
from transformers import Wav2Vec2ForCTC

model = Wav2Vec2ForCTC.from_pretrained(
    "facebook/wav2vec2-base",
    attention_dropout=0.1,
    hidden_dropout=0.1,
    feat_proj_dropout=0.1,
    mask_time_prob=0.05,
    layerdrop=0.1
    ctc_loss_reduction="mean",
    pad_token_id=processor.tokenizer.pad_token_id,
    vocab_size=len(processor.tokenizer),
)

model.freeze_feature_extractor()
```

Figure 2 Hyperparameters for Fine-Tuning Wav2Vec 2.0 Models

Based on Figure 2 above, several hyperparameters were used in the Wav2Vec 2.0 model to enhance its performance and stability during fine-tuning. The dropout rates for attention_dropout, hidden_dropout, and feature_projection were all set to 0.1, aiming to reduce the model's reliance on specific units and prevent overfitting. Additionally, the masking time probability (mask_time_prob) was set to 0.05, meaning that 5% of the input audio would be masked during training, helping the model become more robust to input variations. The LayerDrop setting 0.1 added extra regularization by randomly disabling certain layers during training, which also helped prevent overfitting. Furthermore, the CTC loss reduction (ctc_loss_reduction) was set to "mean" to calculate the average loss value, avoiding bias toward longer samples. The pad_token_id and vocab_size were taken from the tokenizer, ensuring the model could handle all relevant tokens for the task. Additionally, the model used the freeze_feature_extractor function to freeze the initial part of the feature extractor, keeping the feature representations stable so that only the final layers were retrained, thereby reducing computation time and memory usage. This combination of hyperparameters is designed to strike a balance between regularization and flexibility, allowing the model to perform more stable fine-tuning.

2. 5 Evaluation Metrics

The evaluation methods used in this research will be based on Word Error Rate (WER) and Character Error Rate (CER).

Word Error Rate (WER) is an evaluation metric used in speech recognition systems to measure the system's accuracy in recognizing spoken words. WER is calculated by comparing the system's transcription with the reference transcription. It is determined by counting the number of words mispronounced, omitted, or inserted by the system and then dividing that by the total number of words in the reference transcription. The lower the WER value, the better the performance of the speech recognition system. The function of WER is to evaluate the system's performance and indicate how accurately the system recognizes spoken words (Taniya, Bhardwaj, and Kadyan, 2020). WER can be calculated according to Equation (1) below:

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

where:

S is the number of words incorrectly pronounced (substitutions)

I is the number of missing words (insertions)

D is the number of added words (deletions)

N is the total number of words in the test dataset

Character Error Rate (CER) is an evaluation metric used to measure the errors in the transcription text generated by an Automatic Speech Recognition (ASR) system. Compared to the correct reference text, CER is calculated based on the number of character errors, including substitutions, deletions, and insertions.

According to the research conducted by Sawata et al. (2022), CER is calculated by comparing the ASR transcription text with the original reference text. CER is determined by summing all the errors (substitutions, deletions, and insertions of characters) and dividing it by the total number of characters in the reference text. CER can be calculated according to Equation (2) below:

$$CER\% = \frac{S+I+D}{N*100} \quad (2)$$

where:

S is the number of characters incorrectly pronounced (substitutions)

I is the number of missing characters (insertions)

D is the number of added characters (deletions)

N is the total number of characters in the test dataset

3. RESULTS AND DISCUSSION

3. 1 Comparative Analysis of Fine Tuning Model Evaluation Results

The experiments in this study involve fine-tuning the Wav2Vec 2.0 model using two different datasets. The first dataset is the original dataset, and the second is the augmented dataset (a combination of the original training data and the augmented data). Table 1 below shows the comparison of the fine-tuning evaluation results between the original dataset and the augmented dataset. The validation loss for the augmented dataset tends to fluctuate more at

certain steps but remains generally stable overall. The training loss for the augmented dataset decreases more rapidly, indicating that the model learns faster with the augmented data.

Table 1. Comparison of fine-tuning evaluation results.

Epoch	Dataset	Train Loss	Val Loss	WER	CER
100	Original	0.1591	2.32568	0.55249	0.30993
	Augmentation	0.0522	2.68247	0.48619	0.27724
200	Original	0.2024	2.55229	0.54696	0.31719
	Augmentation	0.0177	3.01376	0.45856	0.27119
300	Original	0.0569	2.9414	0.53591	0.3184
	Augmentation	0.0197	3.10229	0.44751	0.27361

For the original dataset, although there was a reduction in WER and CER, the model experienced stagnation after a certain number of steps, even at higher epochs (200 and 300). This indicates that the model could not acquire sufficient learning due to the limited data. Higher warmup steps gave the model more time to adjust to a lower learning rate, but convergence still occurred quickly with limited data. The applied dropout helped reduce overfitting but was not enough to significantly improve WER and CER due to the limited data.

The augmented dataset provided better overall results compared to the original dataset. The faster and more significant reduction in WER and the more stable reduction in CER showed that the model could generalize better due to the larger data. The larger batch size (16) and the increased data allowed the model to train on more examples in each epoch, leading to more efficient training and improved performance. More warmup steps provided sufficient time for the model to adjust the learning rate, while dropout remained effective in reducing overfitting.

3. 2 Analysis of Phoneme Pronunciation Difficulties

In this study, an analysis was conducted on speech recognition results using an Automatic Speech Recognition (ASR) model trained with Wav2Vec 2.0 to identify common pronunciation error patterns among Indonesian children speaking English.

The analysis process began by comparing the ASR-generated transcriptions with verified reference transcriptions. Errors were then categorized based on phoneme types that were frequently mispronounced. These phonemes were analyzed to determine whether the errors occurred consistently in specific positions, such as at the beginning, middle, or end of words.

Additionally, words with a high error rate were identified to examine whether phonological complexity influenced pronunciation difficulty. These words were further analyzed to determine whether specific linguistic factors, such as the presence of consonant clusters or complex vowel combinations, caused the errors.

The error patterns were classified into three main types. The first type is phoneme substitution, where one phoneme is replaced with another, such as /r/ being pronounced as /l/. The second type is phoneme deletion, where a phoneme is omitted, such as the omission of final sounds in certain words. The last type is phoneme addition, where an extra sound is inserted that does not exist in the correct pronunciation.

Table 2. List of phonemes with the highest errors

Phoneme	Word Examples	Error Percentage (%)
/ð/	this, that	78.2
/θ/	three, think	72.5
/r/	red, rabbit	65.3
/v/	very, voice	61.8
/z/	zebra, zoo	58.6
/ʃ/	shoe, sheep	55.9

Based on the phoneme-level error analysis, certain phonemes were found to have significantly higher error rates than others. Table 4.2 lists the phonemes with the highest error rates identified in this study.

The data showed that interdental phonemes /ð/ and /θ/ had the highest error rates, at 78.2% and 72.5%, respectively. This can be attributed to linguistic factors, as these phonemes do not exist in Indonesian. Indonesian children, accustomed to speaking in their native language, struggle to produce these sounds due to the lack of a direct equivalent in the Indonesian phonological system.

In addition to interdental phonemes, /r/, /v/, /z/, and /ʃ/ also exhibited high error rates. The phoneme /r/ can be challenging for some children due to variations in its pronunciation across different Indonesian dialects. Meanwhile, /v/ and /z/ are uncommon in Indonesian vocabulary, leading children to substitute them with more familiar sounds. The phoneme /ʃ/, though present in some loanwords, is still relatively rare in children's daily conversations, contributing to its higher error rate.

3. 3 Analysis of Word Pronunciation Difficulties

Apart from phonemes, an analysis was also conducted on words with the highest pronunciation error rates in the training and testing datasets. These words generally contained difficult phonemes and were often altered during pronunciation. Table 4.3 lists the words with the highest error rates.

Table 3. List of words with the highest errors

Word	Problematic Phonemes	Error Percentage (%)
three	/θ/	72.5
that	/ð/	68.4
rabbit	/r/	65.3
very	/v/	61.8
zebra	/z/	58.6

Most errors occurred because children replaced difficult phonemes with those more familiar in Indonesian. For instance, /ð/ was frequently substituted with /d/, and /θ/ with /t/. This happens because Indonesians lack direct equivalents for these phonemes, leading children to naturally adjust their pronunciation to more familiar sounds.

Additionally, the phoneme /r/ in the word "rabbit" often exhibited pronunciation variations, especially among children who had not yet fully mastered its articulation. The phoneme /v/ in "very" and /z/ in "zebra" were frequently replaced with more common Indonesian sounds, such as /f/ for /v/ and /s/ for /z/. These findings highlight that one of the main challenges in training ASR models for children is addressing frequent phoneme substitutions.

Several approaches can be implemented to improve the ASR model's accuracy in recognizing these words. These include additional training with more diverse pronunciation samples, data augmentation techniques to simulate various pronunciation possibilities, and fine-tuning methods focused on high-error words. By applying these strategies, the model is expected to better recognize children's speech variations and enhance overall performance.

3. 4 Implications for Teaching

This analysis reveals several recommendations for English teachers working with Indonesian children to help improve their pronunciation skills and reduce pronunciation errors.

The first recommendation is Targeted Phoneme Practice. Through phonetic activities and games, teachers can provide exercises specifically designed to target difficult phonemes

such as /ð/, /θ/, and /r/. Techniques such as minimal pair drills, where children distinguish between words with similar phonemes, can help them better understand sound differences.

The second approach is Audio-Based Learning. ASR-based applications can be an effective tool for providing real-time feedback on children's pronunciation. These applications enable children to detect their own mistakes and make corrections independently.

Finally, implementing the Repetitive Pronunciation Model can be beneficial. The listen-and-repeat technique encourages children to imitate native speakers. Audio recordings of native pronunciation or interactive videos can help them become familiar with the correct way to articulate words.

By applying these strategies, children are expected to develop better pronunciation skills and improve their fluency in English communication. Additionally, integrating ASR technology into the learning process can create a more engaging and interactive learning experience for young learners.

4. CONCLUSIONS

This study successfully fine-tuned the ASR Wav2Vec 2.0 model using a dataset of Indonesian children speaking English. The test results indicate that data augmentation significantly improved the model's performance. On the original dataset, the model achieved the lowest Word Error Rate (WER) of 53% at epoch 200 and a final Character Error Rate (CER) of 33%. With the augmented dataset, WER decreased to 45% at epoch 300, while the final CER reached 27%. This improvement of approximately 15% resulted in a more accurate transcription of Indonesian children's speech when pronouncing short English sentences.

An in-depth analysis of Indonesian children's speech recognition using the ASR Wav2Vec 2.0 model identified specific pronunciation errors.

These findings suggest the need for targeted phoneme training, an audio-based approach with ASR feedback, and the listen-and-repeat technique to improve English pronunciation skills among Indonesian children.

ACKNOWLEDGEMENTS

The author would like to thank everyone who contributed to this research, including my parents, who have always supported me, and the teachers of students from the kindergarten and elementary schools who permitted the collection of children’s voice data.

REFERENCES

[1] Aggarwal, C. C. (2018). Neural networks and deep learning. Springer International Publishing AG.

[2] Arisaputra, P., & Zahra, A. (2022). Indonesian Automatic Speech Recognition with XLSR-53. Master of Information Technology, Binus University.

[3] Baevski, A., & Auli, M. (2019). Wav2Vec: Unsupervised pre-training for speech recognition. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 1–10.

[4] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information

processing systems, 33, 12449–12460.

[5] Chen, G., Na, X., Wang, Y., Yan, Z., Zhang, J., Ma, S., & Wang, Y. (2021). Data augmentation for children's speech recognition: The "Ethiopian" system for the SLT 2021 children speech recognition challenge. arXiv preprint arXiv:2010.00171

[6] Banno, S., & Matassoni, M. (2022). Proficiency assessment of L2 spoken English using Wav2Vec 2.0. Paper presented at the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA, 2022), Fondazione Bruno Kessler, Trento, Italy & University of Trento, Trento, Italy.

[7] Bhardwaj, V., Bala, S., Kadyan, V., & Kukreja, V. (2020). Development of robust automatic speech recognition system for children using Kaldi Toolkit. Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020).

[8] Bhardwaj, V., Kadyan, V., & Kukreja, V. (2020). Deep neural network trained Punjabi children speech recognition system using Kaldi Toolkit. Chitkara University Institute of Engineering and Technology.

[9] Chen, L., Asgari, M., & Dodge, H. H. (2023). Optimize Wav2Vec2's architecture for small training sets through analyzing its pre-trained models attention pattern.

[10] Chen, L., Asgari, M., & Dodge, H. H. (2022). Optimize Wav2Vec2's architecture for small training sets through analyzing its pre-trained models attention pattern. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 11.

[11] Dimzon, F. D., & Pascual, R. M. (2009). An automatic phoneme recognizer for children's Filipino read speech. Proceedings of the 2009 International Conference on Machine Learning and Computing, 3, 1–5.

[12] Facebook AI Research. (2020). Wav2Vec 2.0: A framework for self-supervised learning of speech representations.

[13] Jain, R., Barcovschi, A., Yiwere, M., Bigioi, D., Corcoran, P., & Cucu, H. (2023). A Wav2Vec2-based experimental study on self-supervised learning methods to improve child speech recognition. IEEE Access, 11, 30129–30141. https://doi.org/10.1109/ACCESS.2023.3056782.

[14] Jain, R., Yiwere, M., Bigioi, D., Corcoran, P., & Cucu, H. (2017). A Wav2Vec2-based experimental study on self-supervised learning methods to improve child speech recognition. Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Romania, 13.

[15] Kathania, H. K., Kadiri, S. R., Alku, P., & Kurimo, M. (2020). Study of formant modification for children ASR. Department of Signal Processing and Acoustics, Aalto University, Finland.

[16] Moell, B., O'Regan, J., Mehta, S., Kirkland, A., Lameris, H., Gustafsson, J., & Beskow, J. (2022). Speech data augmentation for improving phoneme transcriptions of aphasic speech using Wav2Vec 2.0 for the PSST Challenge. Proceedings of the RaPID-4 @LREC 2022, 62–70.

[17] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779.

[18] Santos, R. L., Caseiro, D., & Trancoso, I. (2018). Mispronunciation detection in children's reading of sentences. Computer Speech & Language, 50, 1–16. https://doi.org/10.1016/j.csl.2018.02.001.

[19] Syahputra, M. E., & Zahra, A. (2021). Unsupervised pre-training pada speech recognition menggunakan bahasa Indonesia berbasis Wav2Vec. Master of Information Technology, Binus University.

[20] Taniya, V., Bhardwaj, V., & Kadyan, V. (2020). Deep neural network trained Punjabi children speech recognition system using Kaldi Toolkit. Chitkara University Institute of Engineering and Technology.

[21] Wills, S., Bai, Y., Tejedor García, C., Cucchiarini, C., & Strik, H. (2021). Automatic speech recognition of non-native child speech for language learning applications. Proceedings of the 2021 Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 38–49. https://doi.org/10.18653/v1/2021.eacl-1.4.

[22] Yu, D., & Deng, L. (2015). Automatic speech recognition: A deep learning approach. Springer-Verlag London. N. Kawasaki, "Parametric study of thermal and chemical nonequilibrium nozzle flow," M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.