Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2022, Article ID 3227828, 16 pages
https://doi.org/10.1155/2022/3227828

Research Article

Acoustic Model with Multiple Lexicon Types for Indonesian Speech Recognition

Taufik Fuadi Abidin,[1] **Alim Misbullah**,[1] **Ridha Ferdhiana**,[2] **Laina Farsiah**,[1] **Muammar Zikri Aksana**,[1] **and Hammam Riza**[3]

[1]*Department of Informatics, Universitas Syiah Kuala, Banda Aceh, Indonesia*
[2]*Department of Statistics, Universitas Syiah Kuala, Banda Aceh, Indonesia*
[3]*National Research and Innovation Agency, Jakarta, Indonesia*

Correspondence should be addressed to Alim Misbullah; misbullah@unsyiah.ac.id

Received 8 May 2022; Accepted 1 September 2022; Published 16 September 2022

Academic Editor: Bhargav Appasani

Currently, speech recognition datasets are increasingly available freely in various languages. However, speech recognition datasets in the Indonesian language are still challenging to obtain. Consequently, research focusing on speech recognition is challenging to carry out. This research creates Indonesian speech recognition datasets from YouTube channels with subtitles by validating all utterances of downloaded audio to improve the data quality. The quality of the dataset was evaluated using a deep neural network. The time delay neural network (TDNN) was used to build the acoustic model by applying the alignment data from the Gaussian mixture model-hidden Markov model (GMM-HMM). Data augmentation was used to increase the number of validated datasets and enhance the performance of the acoustic model. The results show that the acoustic model built using the validated datasets is better than the unvalidated datasets for all types of lexicons. Utilizing the four lexicon types and increasing the data through augmentation to train the acoustic models can lower the word error rate percentage in the GMM-HMM, TDNN factorization (TDNNF), and CNN-TDNNF-augmented models to 40.85%, 24.96%, and 19.03%, respectively.

1. Introduction

The machine's ability to transform voice into text has made it easier in various fields, including simplifying the recording process and displaying video transcripts or subtitles. One advantage is that each person's conversation can be recorded directly and quickly converted into texts [1, 2]. These records can be further indexed for the searching and retrieving processes using keywords [3]. Another benefit of speech-to-text technology is that it allows users to command computers and smartphones through voice [4].

Data show that many research works for speech-to-text in English have been carried out. In [5], Hinton et al. built an acoustic model using the deep neural network (DNN) algorithm to replace the Gaussian mixture model (GMM) technique. Hinton mentioned that a DNN with many hidden layers and nodes could improve accuracy. The work in [6] conducted similar research by changing the DNN structure to long short-term memory (LSTM) to train the acoustic model. In this study, LSTM, which is also part of the recurrent neural network (RNN), can obtain better accuracy when compared to the DNN, although the training process takes more time. In addition, Williams et al. [7] studied language models and built the models using the RNN to optimize the model for a speech-to-text system. The Indonesian speech-to-text studies were conducted by [8–12]. The systems were built using the GMM and hidden Markov model (HMM) techniques and converted the speech spontaneously through dictation. However, the accuracy is still below 85%, and the speech datasets are not publicly available. Hence, it is not easy to compare the methods.

Our contributions are as follows: First, this research builds a speech dataset to assist automatic speech recognition (ASR) voice data processing in Indonesian, the target

Table 1: The methods in the literature, dataset used, and their performance [26–30].

Related work	Method	Dataset	Performance (% WER)
Enhancement of automatic speech recognition by deep neural networks [26]	DNN-HMM, data augmentation	The 34 hours speech of English diverse dataset	16.85%
Self-supervised speech enhancement for Arabic speech recognition in real-world environment [27]	Denoising auto encoder, HMM	The Arabic mobile parallel speech multi-dialect speech corpus	30.17%
Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions [28]	Pitch enhancement, DNN-HMM	The Punjabi adult/child speech dataset	10.98%~12.24%
A hybrid speech enhancement algorithm for voice assistance application [29]	Noise suppression, HMM	The 8.5 hours English medical speech dataset (RAVDESS)	17.5%~22.9%
Dual application of speech enhancement for automatic speech recognition [30]	RNN transducer, data augmentation	The social media English video dataset	8.3%~13.4%

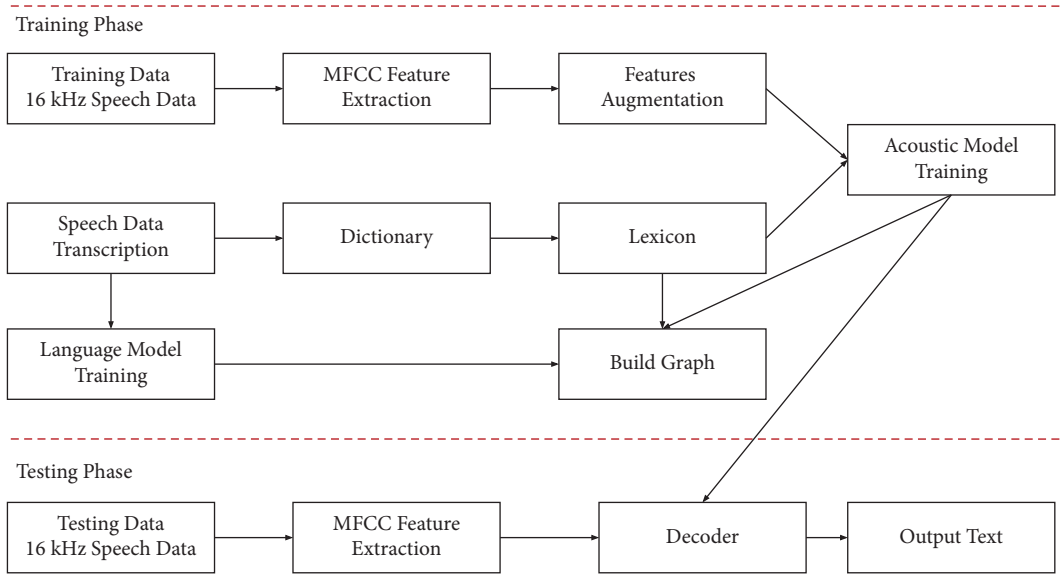

Figure 1: The illustration of creating the Indonesian speech recognition system.

language. The data sources are the videos from YouTube with subtitles [13]. The YouTube channel manages many videos and provides a feature to download them. The video on YouTube channels generally has transcripts, a series of writings placed at the bottom of the video related to the conversation. Voice and video transcripts are downloaded to build a speech-to-text dataset. Second, this research provides an open Indonesian speech-to-text dataset. Third, this research builds acoustic models by applying the alignment data from the Gaussian mixture model-hidden Markov model (GMM-HMM), TDNN factorization (TDNNF), and CNN-TDNNF-augmented models. Fourth, data augmentation is utilized to increase the number of validated datasets and improve the performance of the acoustic model.

1.1. Related Works. Research related to speech-to-text has been widely carried out in various languages such as English, Mandarin [14, 15], African [16], Pakistani [17], Italian [18], and Indian [19]. Most works built acoustic models using the Mel frequency cepstral coefficient (MFCC) feature, a stable and accurate cepstral coefficient representing sound and music [20, 21].

In addition, many acoustic speech-to-text models were built using a DNN that utilizes the alignment model of the GMM-HMM as the target class to train the model [22, 23]. Various DNNs have been developed to build acoustic models. The work in [23] used the DNN with four hidden layers with 1.500 nodes for each layer to build the acoustic model from heterogeneous datasets. They found a phoneme error rate of 12.76% for children, 10.91% for adult women, and 8.62% for adult males. In [24], Sak used the LSTM-RNN architecture to train an acoustic model from a Google voice search dataset of three million utterances or approximately 1.900 hours. The word error rate (WER) was 10.7% for test data containing 22,500 utterances. The research by Zia and Zahid [25] also used the same architecture to train an acoustic model for the Urdu language dataset of 20 speakers consisting of hundreds of words. The acoustic model was trained using several types of LSTM-RNN architecture, such

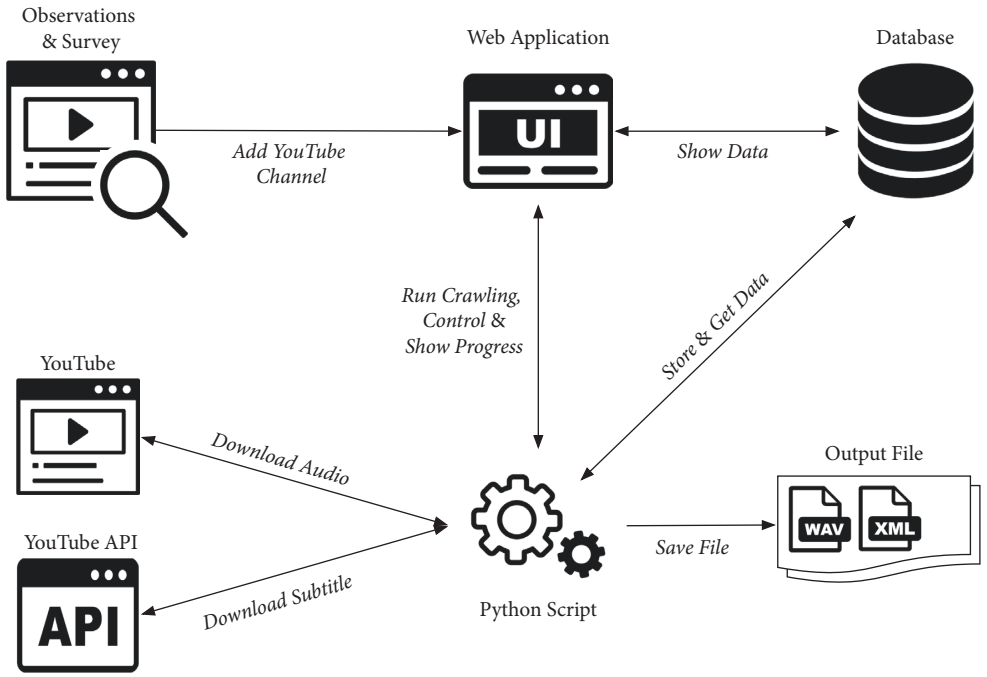

Figure 2: The process of collecting Indonesian audio from the YouTube channel.

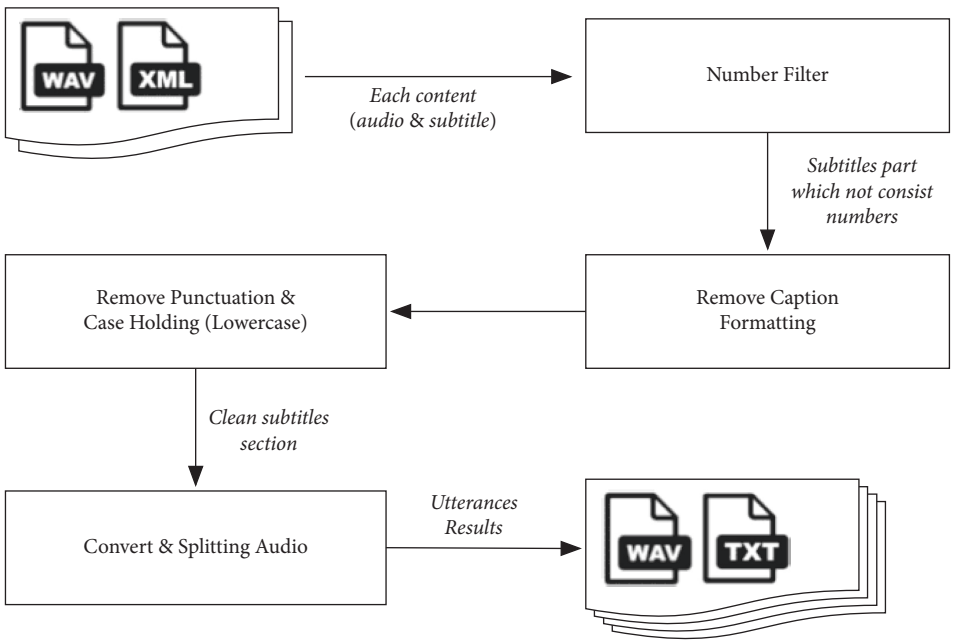

Figure 3: The data preparation process.

Table 2: The format of Kaldi ASR toolkit.

Filename	Format
Text	z-xWY3INfpBGo_56. The most important thing is that we keep moving forward z-ztfq4QIuuDU_281 increase knowledge
wav.scp	z-xWY3INfpBGo_56~/path-to-file/z-xWY3INfpBGo_56.wav z-ztfq4QIuuDU_281~/path-to-file/z-ztfq4QIuuDU_281.wav
utt2spk	z-xWY3INfpBGo_56 z-xWY3INfpBGo_56 z-ztfq4QIuuDU_281 z-ztfq4QIuuDU_281
spk2utt	z-xWY3INfpBGo_56 z-xWY3INfpBGo_56 z-ztfq4QIuuDU_281 z-ztfq4QIuuDU_281

Table 3: The Indonesian audio dataset before validation.

Category	Original audio		Utterances	
	Number of audio	Total duration (hours)	Number of utterances	Total duration (hours)
Autos and vehicles	20	2.8039	2,234	1.3289
Comedy	40	7.7125	6,017	4.0768
Education	553	54.9508	47,477	38.4886
Entertainment	308	57.1747	39,155	25.2972
Film and animation	77	7.7464	6, 475	4.5591
Gaming	1	0.1150	124	0.0987
Howto and style	355	54.5117	28,726	41.8189
Music	56	4.6469	1, 409	2.5783
News and politics	170	24.6003	19,295	16.8386
People and blogs	179	50.2578	32,205	22.7193
Pets and animals	2	0.3153	57	0.0433
Science and technology	215	44.5308	31,492	22.9787
Sports	1	0.1583	137	0.1131
Travel and events	1	0.2236	78	0.0633
Uncategorized	2	0.3367	410	0.2515
Total	1,980	310.0847	215,291	181.2543

Table 4: The dataset of Indonesian audio after first validation (approx. size 10,000 utterances).

Category	Utterances before validation		Utterances after validation	
	Number of utterances	Total duration (hours)	Number of utterances	Total duration (hours)
Autos and vehicles	2,234	1.3289	29	0.0214
Comedy	6,017	4.0768	71	0.0498
Education	47,477	38.4886	5,973	4.5254
Entertainment	39,155	25.2972	470	0.2994
Film and animation	6,475	4.5591	75	0.0567
Gaming	124	0.0987	1	0.0000
Howto and style	28,726	41.8189	450	0.5933
Music	1,409	2.5783	2	0.0011
News and politics	19,295	16.8386	303	0.2502
People and blogs	32,205	22.7193	2,432	1.7630
Pets and animals	57	0.0433	1	0.0000
Science and technology	31,492	22.9787	515	0.3849
Sports	137	0.1131	1	0.0000
Travel and events	78	0.0633	1	0.0008
Uncategorized	410	0.2515	9	0.0069
Total	215,291	181.2543	10,333	7.9529

as plain, deep architecture, bidirectional, and deep-directional. The architecture achieved a 15% performance increase compared to the baseline architecture.

Other architectures were also used to build acoustic models to study long-term temporal relationships from speech data, including the time delay neural network (TDNN) [31, 32]. Meanwhile, convolutional neural networks (CNNs) [22, 33, 34] were used to improve the performance compared to the previous architecture when using English datasets Table 1.

Various studies were recently conducted to reduce word error rates using different techniques. One of the standard techniques is data augmentation (DA), such as the work done by [26] using the 34 hours of speech from a diverse English dataset. In [26], the performance of speech recognition was successfully reduced to 16.85% by implementing data augmentation using the DNN. Similarly, the DNN combined with pitch enhancement was also used by [28] on the Punjabi adult and children's speech dataset to build different acoustic model conditions. Bhardwaj and Kukreja [28] enhanced the pitch using cepstral analysis in the feature extraction process and achieved a 10.98%~12.24% WER under different acoustic conditions.

Speech-to-text for the Indonesian language has been conducted using various Indonesian speech recognition datasets [8–12, 35, 36]. Winursito [9] used principal component analysis (PCA) to reduce the dimension size of the MFCC features from 26 to 10. The accuracy of the speech-to-text system increased from 86.43% to 89.29%, trained with several speakers with 28 and 140 utterances. Another study conducted by Teduh [11] built a corpus of speech in Indonesian with 100,000 utterances recorded by 400 speakers from various regions in Indonesia. The corpus was used to build a speech-to-text model with a 20% WER.

Our work differs from the previous studies mentioned above in terms of the spontaneous speech dataset which was collected from YouTube, different lexicon constructions, and acoustic model construction using validated and unvalidated

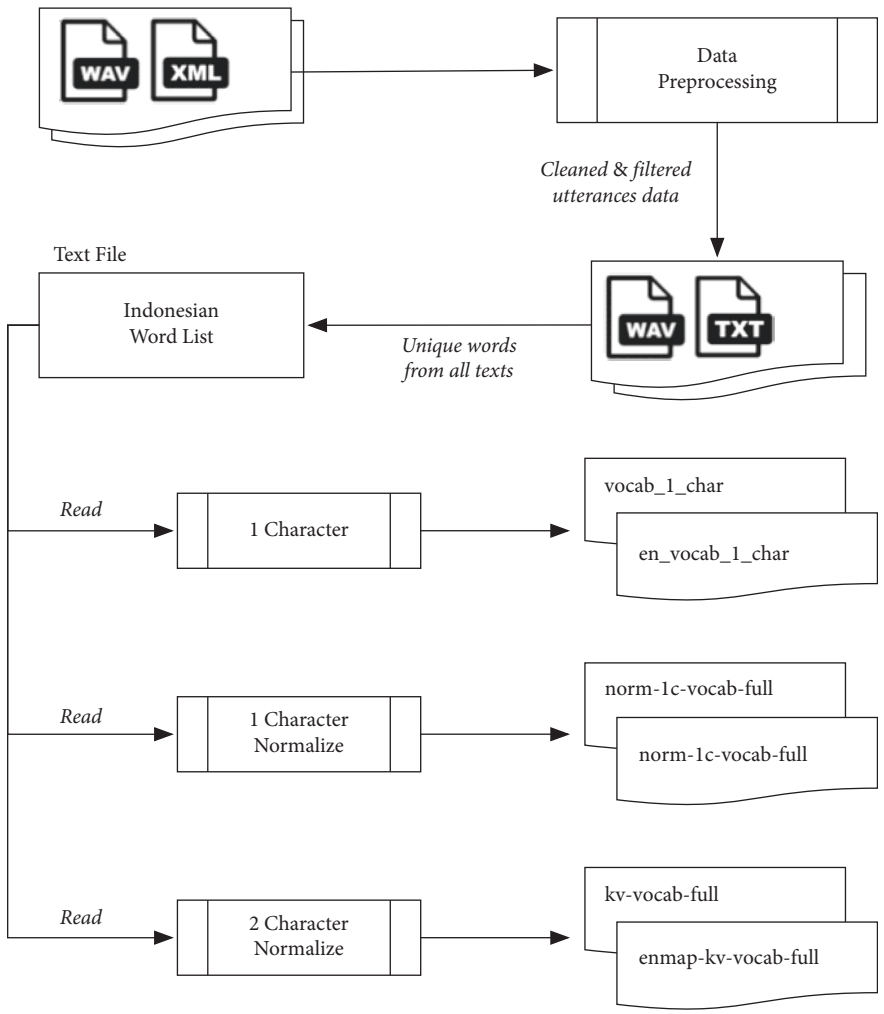

Figure 4: The extraction process of the lexicon from a dictionary.

Table 5: Total pronunciations in the lexicons.

Lexicon type	Total pronunciation
vocab_1_char	26
enmap_vocab_1_char	49
norm_1c_vocab_full	26
enmap_norm_1c_vocab_full	49
kv_vocab_full	135
enmap_kv_vocab_full	157

datasets. As for the datasets, we collected audio from the YouTube channels and built the acoustic models using GMM-HMM, TDNN, and CNN.

2. Materials and Methods

The construction of the speech-to-text system consists of several stages. It begins with collecting audio data from a YouTube channel with a transcript so that the tokenization process of the audio that represents the existing transcript becomes easier. At this stage, checking for fake subtitles is also carried out. If the audio is detected as having a transcript, but the transcript's length does not meet the specified threshold, then the transcript is inaccurate and will not be downloaded.

The second stage is the audio tokenization process, which is based on the transcript at a specific duration. Each sentence in the transcript is extracted based on the start time and duration. At this stage, detecting empty transcripts but having a duration and start time is essential to remove them from the dataset. Third, the audios extracted into utterances and transcripts were cleaned by removing punctuation marks, changing uppercase to lowercase letters, and removing meaningless symbols.

Next, the fourth stage is to build a speech-to-text system through learning and testing, as illustrated in Figure 1. Before training and testing, the unique words in the transcript are extracted for the training. Furthermore, pronunciation is built for each unique word. Lexicon is used as a label in the acoustic model training process using the DNN. In learning the acoustic model, the characteristics of the voice data are extracted using the MFCC method. The data is a matrix representing the speech information in the audio

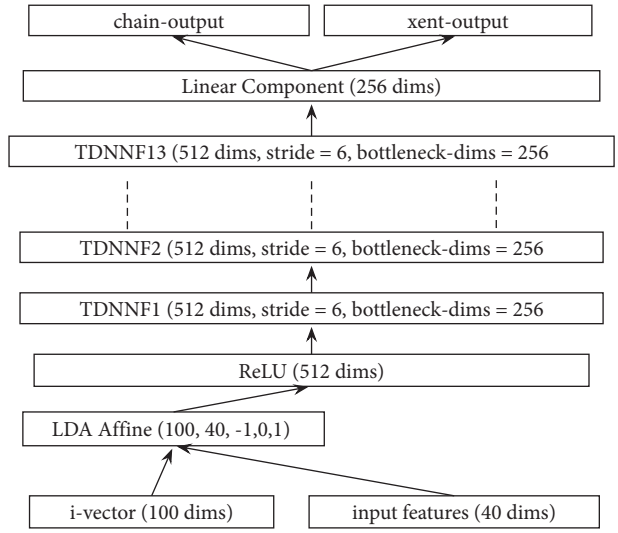

Figure 5: DNN structure for the acoustic model [39].

Table 6: The performance of the acoustic models using the unvalidated dataset.

Lexicon type	GMM-HMM model (%WER)				TDNNF (% WER)
	Monophone	Triphone DELTA + DELTA-DELTA	Triphone LDA + MLLT<	Triphone SAT	
vocab_1_char	76.08	56.06	46.09	47.05	29.62
enmap_vocab_1_char	75.93	54.92	46.15	47.04	29.77
norm_1c_vocab_full	76.18	56.08	46.56	46.36	29.89
enmap_norm_1c_vocab_full	76.87	55.57	46.25	46.17	29.79
kv_vocab_full	71.19	58.61	47.82	48.81	29.49
enmap_kv_vocab_full	72.05	58.55	48.29	47.64	**29.41**

Table 7: Three different sizes of training utterances.

Dataset	Number of utterances		
Training	10,000	30,000	46,550
Testing	2,450	2,450	2,450

data at each frame. After the extraction process, the features are augmented to increase the training data size.

The last stage is acoustic model testing (decoding). Before evaluating the acoustic model, the language model for constructing the graph must be built. The graph contains a lexicon used to train the acoustic model and its weighted equivalents in the language model. The graph is used to find the right word equivalent based on the acoustic model's probability value of the syllables. In the end, the output is in the form of text based on the speech data used. The performance of the model is measured using WER as follows:

$$\text{WER} = \frac{(S + D + I)}{N} = \frac{(S + D + I)}{(S + D + C)}, \quad (1)$$

where S is the number of words replaced, D is the number of words deleted, I is the number of words added, C is the number of correct words, and N is the total number of words in the transcript reference.

2.1. Data Collection and Preparation. We downloaded the audio data in the Indonesian language from the YouTube channel with a specific duration based on the transcript. All audios collected were transcribed and used as a candidate dataset for the Indonesian speech-to-text system. The steps are shown in Figure 2. During data acquisition, YouTube IDs containing speech and transcripts were stored in a database. The audio and its transcript were downloaded using the YouTube API. Each downloaded file was saved in WAV and XML formats. The audio

Table 8: The performance of the acoustic models using 10,000 validated utterances.

Lexicon type	GMM-HMM model (%WER)			
	Monophone	Triphone DELTA + DELTA-DELTA	Triphone LDA + MLLT	Triphone SAT
1c-vocab-full	82.30	79.61	62.80	**63.20**
enmap-1c-vocab-full	82.47	81.05	64.25	64.54
enmap-kv-vocab-full	80.93	86.86	76.61	75.33
kv-vocab-full	79.27	88.00	77.36	76.47

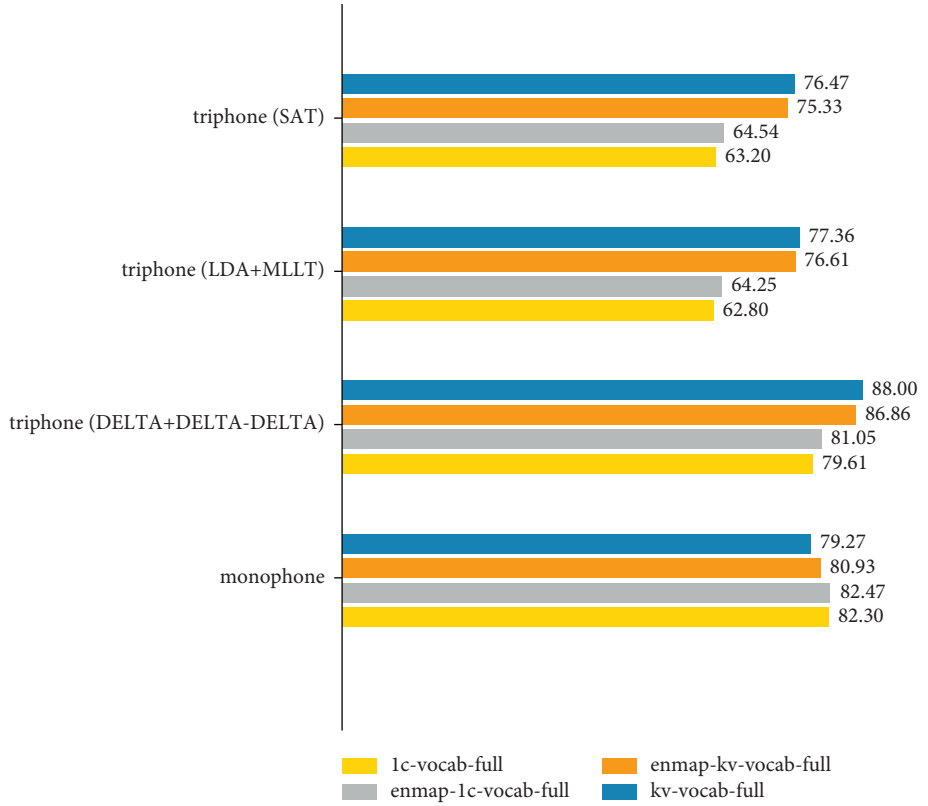

Figure 6: The %WER of the GMM-HMM models trained using 10,000 validated utterances.

Table 9: The performance of the acoustic models using 10,000 unvalidated utterances.

Lexicon type	GMM-HMM model (%WER)			
	Monophone	Triphone DELTA + DELTA-DELTA	Triphone LDA + MLLT	Triphone SAT
1c-vocab-full	85.61	85.26	71.30	**71.24**
enmap-1c-vocab-full	86.83	86.12	73.06	73.07
enmap-kv-vocab-full	85.30	91.16	84.28	84.06
kv-vocab-full	84.12	91.11	85.28	85.69

sampling rate was lowered from 44 kHz to 16 kHz, and the channel was changed from stereotype to monotype. Last, the audio was cut based on each transcript's duration information.

The subsequent process is data preparation, i.e., cleaning each transcript and matching the audio and the transcript, as shown in Figure 3. The transcript containing numbers was eliminated because it is complicated to produce the pronunciation.

In this study, the Kaldi ASR toolkit [37] was used to read a unique file format to train the acoustic model, as shown in Table 2. The text file contains the ID and transcript for each audio, while the wav.scp contains the id and the location where the audio was stored. Meanwhile, the utt2pk and spk2utt files contain the ID of each speech and its speaker. There were 1,980 audio data with transcripts collected from the YouTube channels, grouped into 15 categories, as summarized in Table 3. The total duration of all transcripts

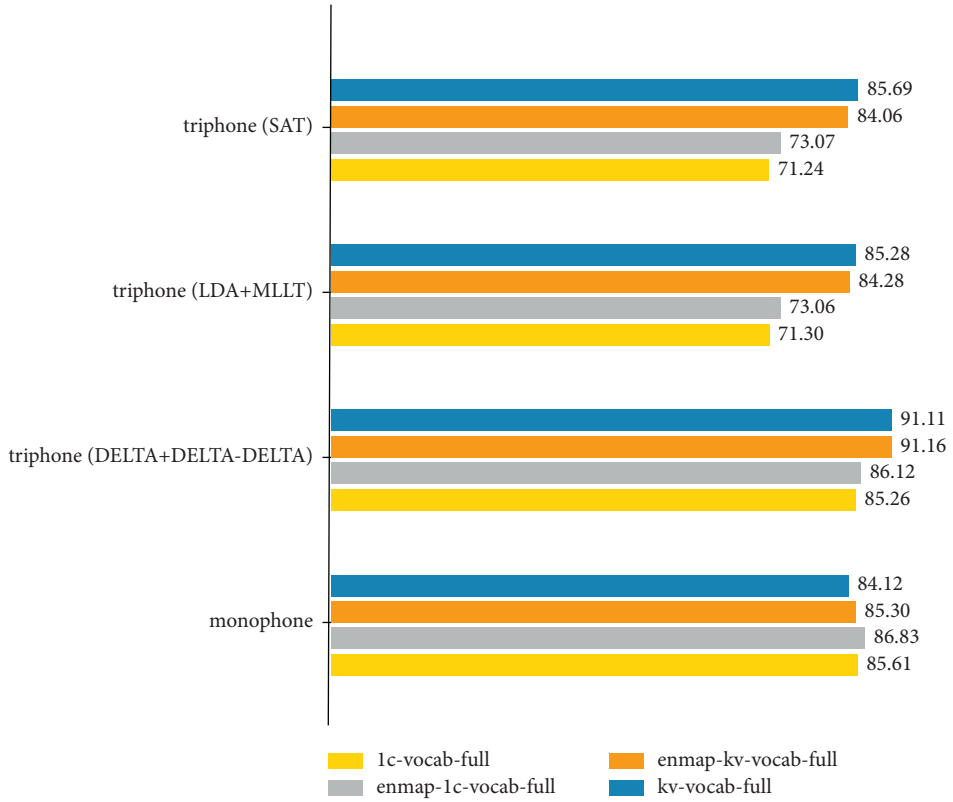

Figure 7: The %WER of the GMM-HMM models trained using 10,000 unvalidated utterances.

Table 10: The performance of the acoustic models using 30,000 validated utterances.

Lexicon type	GMM-HMM model (%WER)			
	Monophone	Triphone DELTA + DELTA-DELTA	Triphone LDA + MLLT	Triphone SAT
1c-vocab-full	81.75	65.41	54.13	**54.90**
enmap-1c-vocab-full	81.34	65.23	54.00	55.09
enmap-kv-vocab-full	80.76	71.21	61.66	62.55
kv-vocab-full	77.16	69.85	60.76	61.73

before cleaning is 310.09 hours. After cleaning, it is shortened to 181.25 hours.

The validation process was carried out for each utterance to get the best speech quality from the audio. The utterances were validated by several validators using a simple validation interface created for the process. The total number of validated utterances in the first round was 10,333, with a 7.953 hour duration, as summarized in Table 4. The validated dataset contains all utterances that have been validated by the validators, whereas the unvalidated dataset contains the original utterances without the validation process.

Before training the acoustic model, several types of vocabulary (lexicon) were prepared. The lexicon was extracted from the unique words in the transcripts. The process of generating the lexicon is shown in Figure 4. About 41,351 unique words in the dictionary were successfully extracted from the transcript, consisting of 3,330 English and 6,334 Indonesian words. The rest are informal slang words.

There are four lexicon types used in this work. Each type is described in detail in [38]. The words in the lexicon are the mapping words to syllables or their pronunciations. The pronunciations are used for different words so that the number of pronunciations is not greater than the number of words in the dictionary, as summarized in Table 5. Lexicon vocab_1_char and norm_1_vocab_full have 26 pronunciations extracted as a character. In contrast, enmap_vocab_1_char and enmap_norm_1_vocab_full have 49 different pronunciations because each word is mapped to the CMUDict English dictionary. Furthermore, kv_vocab_full has 135 pronunciations summed up into 26 single characters, 21 consonants multiplied by 5 vowels, and 4 consonants combinations that make up consonants/kh/,/ng/,/ny/, and/sy/.

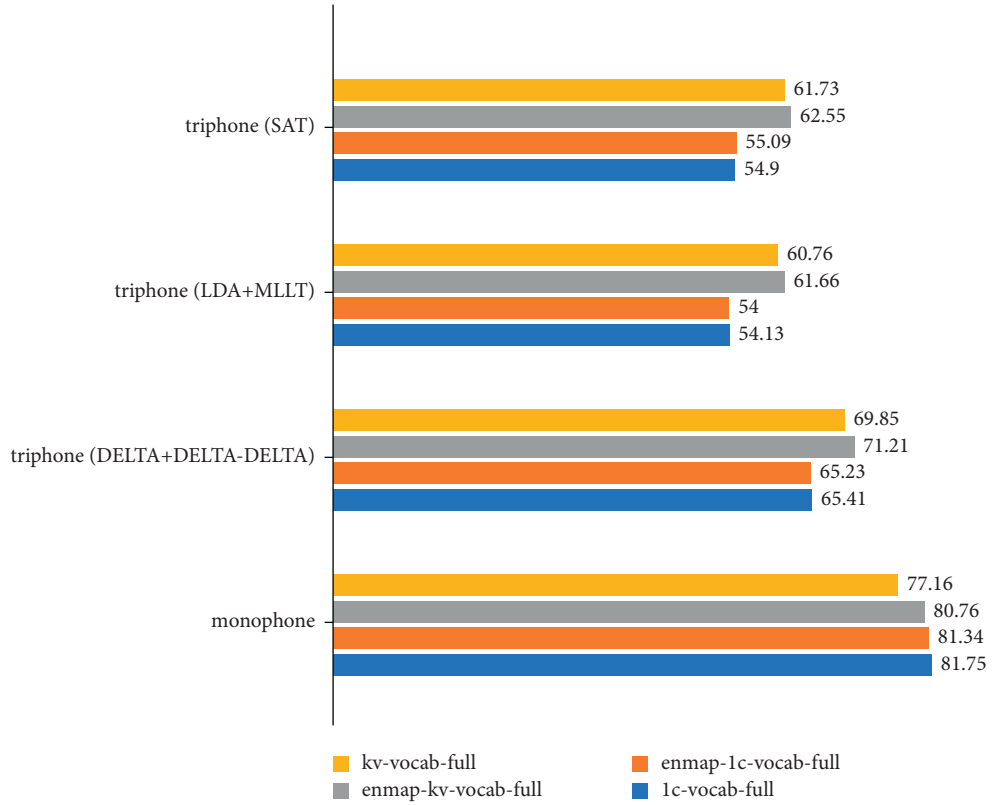

Figure 8: The %WER of the GMM-HMM models trained using 30,000 validated utterances.

Table 11: The performance of the acoustic models using 30,000 unvalidated utterances.

Lexicon type	GMM-HMM model (%WER)			
	Monophone	Triphone DELTA + DELTA-DELTA	Triphone LDA + MLLT	Triphone SAT
1c-vocab-full	85.59	72.34	63.44	**63.28**
enmap-1c-vocab-full	85.34	72.35	64.33	64.46
enmap-kv-vocab-full	84.09	77.01	70.14	70.54
kv-vocab-full	82.15	77.36	70.56	71.00

2.2. Experimental Setup. In this study, we built several acoustic and language models. We trained the acoustic model using the Kaldi ASR toolkit. The MFCC feature was extracted with 40 dimensions and a window size for each frame of 25 ms, adding a shift of 10 ms to the audio duration. The feature was extracted for each audio track. In addition, the acoustic models were also trained using the GMM-HMM technique for monophone and triphone with different techniques such as DELTA, +DELTA-DELTA, and speaker adaptation training (SAT) features. The GMM-HMM acoustic model aligns the training data to get each frame's appropriate labels (phonemes). The matched training data are later used to train the model using the DNN.

The time delay neural network factorization (TDNNF) [39] is used in this study. The idea is to decompose the existing TDNN structure into a small matrix whose dimensions can be multiplied. For example, a TDNN structure has a hidden layer with 700 dimensions. The structure's weights (parameters) are a matrix of size $700 \times 2{,}100$, where 2,100 is obtained from 3 frames consisting of the number of frames and their right and left offsets multiplied by the dimensions of the hidden layers. The $700 \times 2{,}100$ matrix is factorized into $2M = AB$ matrices with 250 dimensions; then, the A matrix size becomes 700×250, and the B matrix size becomes $250 \times 2{,}100$, with the B matrix being semi-orthogonal. In the illustration above, the value 250 is the linear bottleneck dimension, and 700 is the hidden layer dimension. The number of hidden layers used in the TDNN structure is 13, with 512 dimensions for each layer and 80 for linear bottlenecks, while the steps (stride) for each frame are 6. The TDNN structure [39] is shown in Figure 5.

The Lattice-Free Maximum Mutual Information (LF-MMI) was adopted as the objective function and the chain model in the Kaldi ASR toolkit to train the acoustic model using the TDNN structure. In Figure 5, the TDNN structure

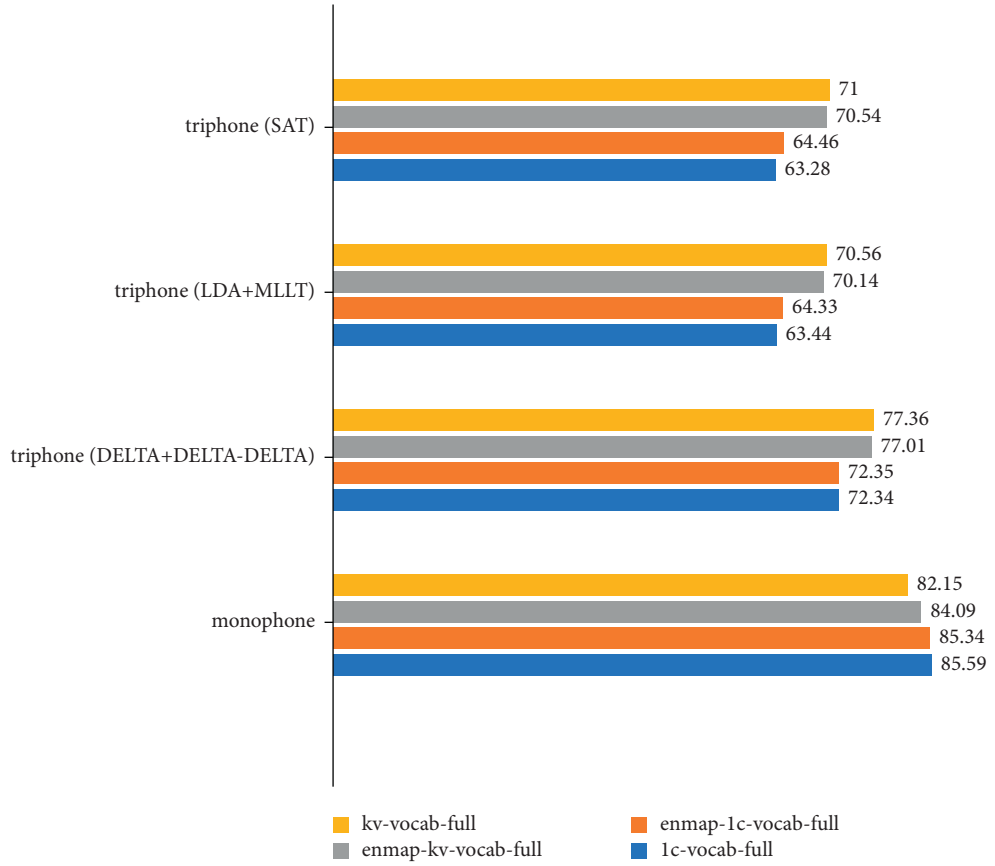

Figure 9: The %WER comparison for different GMM-HMM models using 30,000 unvalidated utterances.

Table 12: The performance of the acoustic models using 46,550 validated utterances.

Lexicon type	GMM-HMM model (%WER)			
	Monophone	Triphone DELTA + DELTA-DELTA	Triphone LDA + MLLT	Triphone SAT
1c-vocab-full	81.25	61.96	50.66	51.24
enmap-1c-vocab-full	80.99	60.88	50.08	50.83
enmap-kv-vocab-full	79.46	65.92	55.23	56.05
kv-vocab-full	76.75	65.27	55.52	56.05

receives input in 40-dimensional features and an i-vector with 100 dimensions. Both are then combined. Furthermore, the dimension reduction is carried out using the latent Dirichlet analysis (LDA) technique before being sent to the hidden layer. The acoustic model is trained for different lexicons and uses five epochs with initial and final learning rates of 0.0015 and 0.00015, respectively. Furthermore, the language model was trained using the SRILM toolkit with 3 grams, which was used to evaluate the acoustic model. The language model was trained using the transcript of the training data.

3. Results and Discussion

As described previously, the acoustic model was built using several lexicon types. The acoustic model was trained using unvalidated and validated datasets. The acoustic models were built using unvalidated utterances. The number of unvalidated utterances used to train and test the models was 206,206 and 5,287, respectively. The best acoustic model utilized the enmap_kv_vocab_full lexicon type, with a WER of 29.41%, trained with TDNNF, as summarized in Table 6.

The subsequent evaluations were carried out using three different sizes of validated utterances. The number of validated utterances is 49,000, which later will be evaluated using several lexicon types. The training utterances for evaluation are summarized in Table 7.

The lexicon used are 1c-vocab-full; enmap-1c-vocab-full; enmap-kv-vocab-full; and kv-vocab-full. It was found that the two types of the lexicon (norm-1c-vocab-full and enmap-norm-1c-vocab-full) used in the previous test were very similar to the lexicon types 1c-vocab-full and enmap-1c-vocab-full, so that these two types of the lexicon were not used in this evaluation. The tests were carried out for both validated and unvalidated utterances.

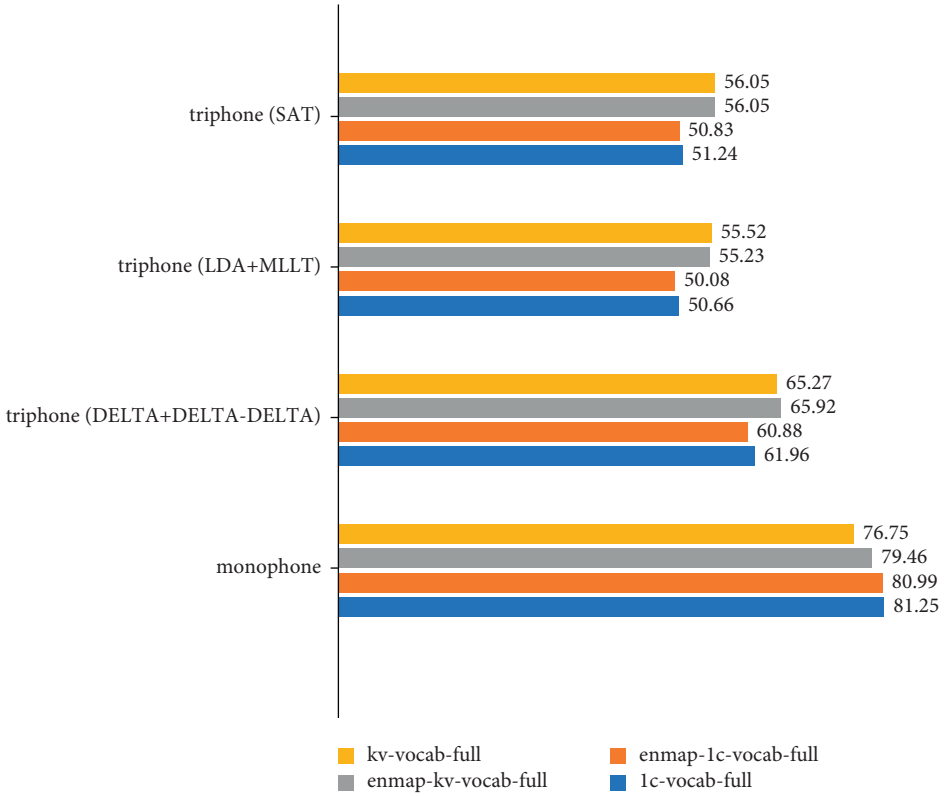

Figure 10: The %WER comparison for different GMM-HMM models using 46,550 validated utterances.

Table 13: The performance of the acoustic models using 46,550 unvalidated utterances.

Lexicon type	GMM-HMM model (%WER)			
	Monophone	Triphone DELTA + DELTA-DELTA	Triphone LDA + MLLT	Triphone SAT
1c-vocab-full	85.77	69.21	60.44	60.45
enmap-1c-vocab-full	86.10	69.65	61.40	61.10
enmap-kv-vocab-full	83.58	73.46	65.22	65.19
kv-vocab-full	82.56	73.38	65.98	65.96

The following evaluations were conducted by training the GMM-HMM models using 10,000 validated utterances. The model's performance was evaluated using 2,450 validated testing utterances, summarized in Table 8. The results show that the smallest WER is 63.20%, generated by the GMM-HMM triphone (SAT) model, trained using the 1c-vocab-full lexicon type and 10,000 validated utterances. Figure 6 compares the %WER of GMM-HMM models.

Furthermore, the following evaluation was also carried out using 10,000 unvalidated utterances. The results, summarized in Table 9, show that the lowest WER percentage for the GMM-HMM model is 71.24%. The performance was obtained using the lexicon type of 1c-vocab-full. However, the WER percentage is greater than the WER percentage of the acoustic model trained using 10,000 validated utterances. Figure 7 compares the %WER of GMM-HMM models using 10,000 unvalidated utterances. The GMM-HMM model trained using the validated utterances outperformed the model trained using the unvalidated utterances.

The following evaluations were conducted for the GMM-HMM models trained using 30,000 validated utterances. The model's performance was tested using the same 2,450 validated testing utterances. The best GMM-HMM model was obtained by the 1c-vocab-full lexicon type with a percentage of WER of 54.9%. The percentage is better than the GMM-HMM model on 10,000 validated data, i.e., 63.2%. These results illustrate that increasing the number of validated data can improve the model's performance and reduce %WER. The results are summarized in Table 10. Figure 8 compares the %WER on several GMM-HMM models using 30,000 validated utterances.

The evaluation was also done for the GMM-HMM models trained using 30,000 unvalidated utterances. We found that the %WER of the GMM-HMM models was even worse than the model trained using the validated utterances, i.e., 63.28%, obtained using the lexicon type of 1c-vocab-full

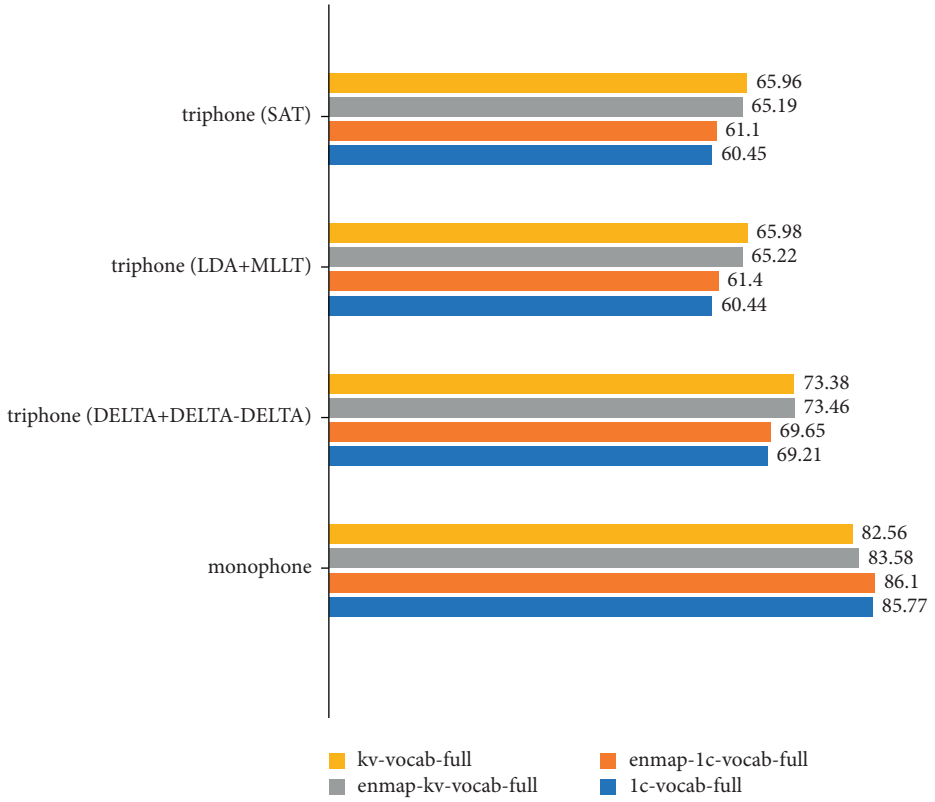

Figure 11: The %WER comparison for different GMM-HMM models using 46,550 unvalidated utterances.

Table 14: The performance of the models using 46,550 validated utterances, for the four-combination lexicon type.

Lexicon type	GMM-HMM model (%WER)			
	Monophone	Triphone DELTA + DELTA-DELTA	Triphone LDA + MLLT	Triphone (SAT)
1c-vocab-full	81.25	61.96	50.66	51.24
enmap-1c-vocab-full	80.99	60.88	50.08	50.83
enmap-kv-vocab-full	79.46	65.92	55.23	56.05
kv-vocab-full	76.75	65.27	55.52	56.05
4-combination	72.17	51.27	41.57	**42.80**

as summarized in Table 11. The %WER was close to the %WER of the model trained using 10,000 unvalidated utterances. This result further shows that increasing the number of validated utterances can improve the model's performance and reduce %WER. Figure 9 compares the %WER of GMM-HMM models using 30,000 unvalidated utterances. Again, we discovered that the GMM-HMM model trained using the validated utterances outperformed the one trained using the unvalidated utterances.

We also trained the GMM-HMM models using 46,550 validated utterances and tested the models using 2,450 validated utterances. We found that the lowest %WER of the models was obtained using the enmap-1c-vocab-full lexicon type, i.e., 50.83%, as shown in Table 12. Moreover, Figure 10 compares the %WER of the models using 46,550 validated utterances.

When we trained the GMM-HMM acoustic models using 46,550 unvalidated utterances, the %WER of the models was not even better. The best %WER was 60.45%, worse than %WER of the models that were trained using the validated utterances, as summarized in Table 13. Figure 11 compares the %WER of GMM-HMM models using 46,550 unvalidated utterances.

All lexicon types were combined in the subsequent evaluation. We named it the four-combination lexicon. The lexicon combines several ways of pronouncing the four lexicons into one lexicon. This combination aims to obtain different pronunciation information that can be trained with a single model. Table 14 shows that %WER for the four-combination lexicon is lower than the %WER of the other GMM-HMM acoustic models, i.e., 42.8%. Figure 12

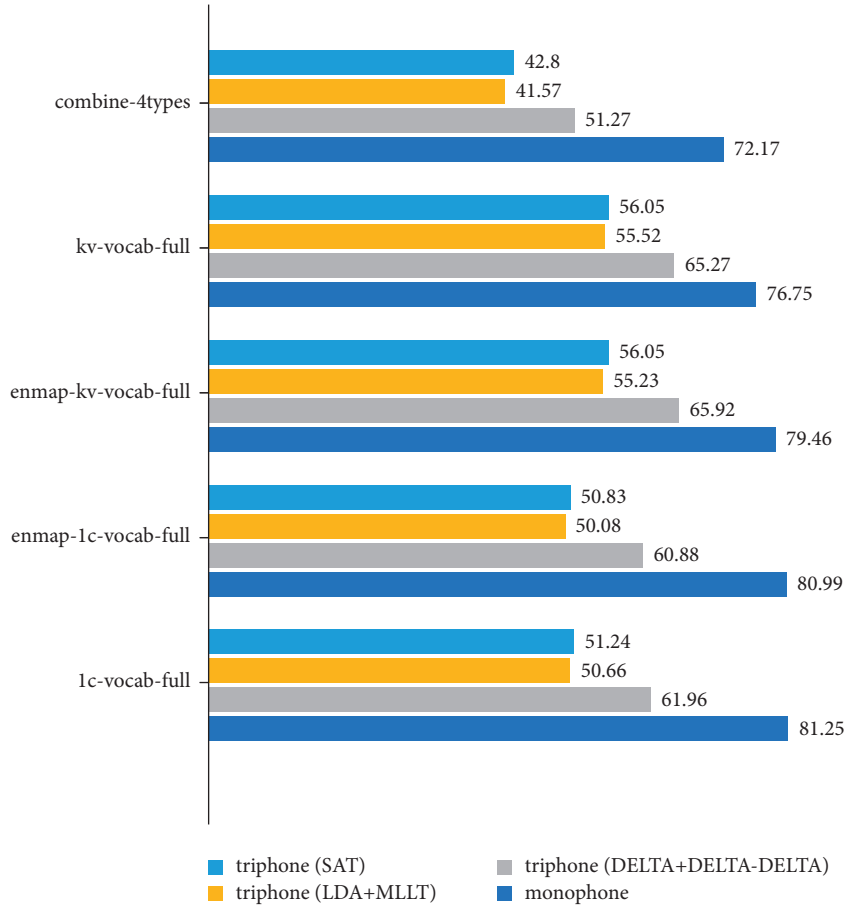

Figure 12: The %WER comparison for different lexicon types on the GMM-HMM models.

Table 15: The performance of the models using 46,550 validated augmented vs. nonaugmented utterances.

Data type	GMM-HMM and TDNNF model (%WER)				
	Monophone	Triphone DELTA + DELTA-DELTA	Triphone LDA + MLLT	Triphone SAT	TDNNF
Validated utterances	72.17	51.27	41.57	42.80	25.10
Validated augmented utterances	73.80	49.52	39.83	**40.85**	**24.96**

compares the %WER of GMM-HMM models using 46,550 validated utterances.

In the following evaluation, we increased the amount of data using the data augmentation approach by changing the tempo of the original audio. The factor values are 0.9 to slow down and 1.1 to speed up. The size of the augmented data is three times larger than the nonaugmented data. The results show that %WER of the GMM-HMM acoustic model decreased when trained using the augmented utterances, i.e., 40.85%. In addition, the %WER of the acoustic model trained using TDNNF was also the smallest when the augmented utterances were utilized, as summarized in Table 15. These results confirmed that %WER decreased when more validated utterances were used. Figure 13 compares the %WER of GMM-HMM models and TDNNF even if data training was augmented or not.

Furthermore, the CNN model was trained using validated augmented utterances and the 4-combination lexicon type. The result shows that the acoustic model returns the best %WER, i.e., 19.03% using the same testing utterances. Table 16 shows the result.

It is not easy to compare the WER of the models for other languages with ours because the datasets are different. Moreover, most datasets are not spontaneous speech datasets like ours, collected from YouTube, which is very noisy and sometimes contains overlapped voices. If the WER of our best result is compared with the one conducted by [8] on their spontaneous speech dataset, our WER, 19.03%, is considerably lower than theirs, i.e., 43.14%. This is also true for the work in [11]. They used speech corpora from several recording projects, which is not a spontaneous speech dataset, and their WER is 20%, slightly higher than ours.

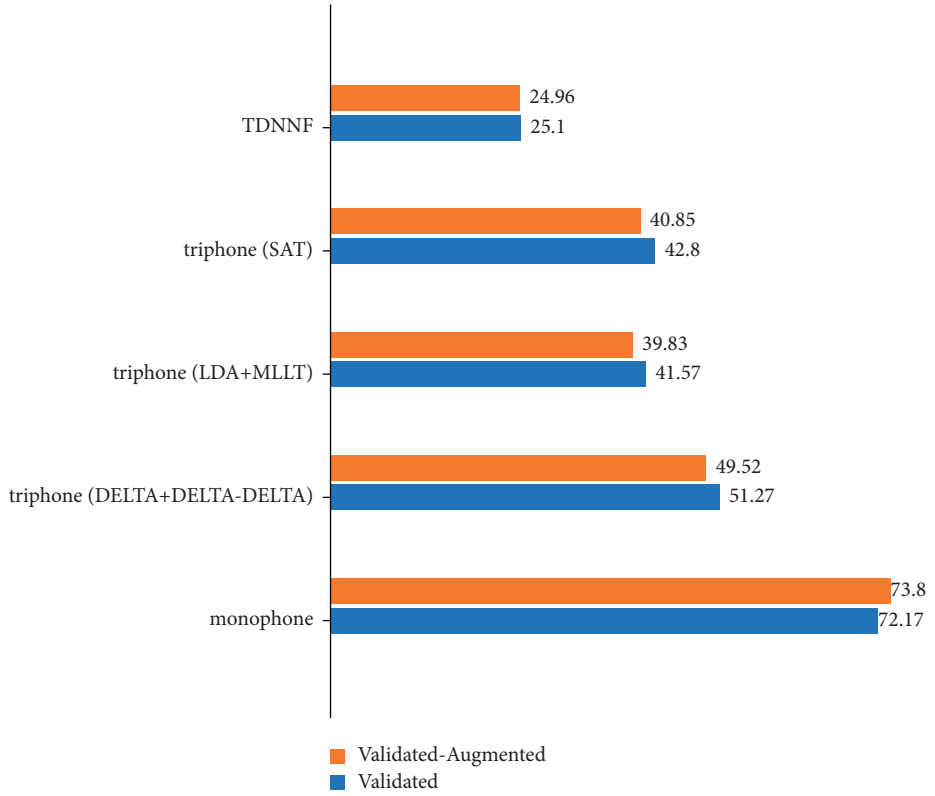

Figure 13: The %WER using 46,550 validated augmented vs. nonaugmented utterances.

Table 16: The performance of the models.

	WER (%)	
Lexicon type	TDNNF	CNN-TDNNF-augmented
4-combination	24.96	**19.03**

4. Conclusions

The 181 hours audio dataset was collected from the YouTube channel, consisting of 215,291 utterances with transcripts. A dictionary of 41,351 unique words was also extracted from the transcripts and used to construct the four lexicon types with different pronunciation patterns. The validated dataset and all lexicon types were used to train the acoustic models with a TDNN approach. The results show that the acoustic model built using the validated dataset is better than the one trained using the unvalidated dataset for all lexicon types. When the acoustic models were trained using the combination of all lexicon types and augmented utterances, the % WER of the GMM-HMM, TDNNF, and CNN-TDNNF-augmented models was reduced to 40.85%, 24.96%, and 19.03%, respectively.

The limitation of this work is that the size of the validated utterances was considered small to recognize all phonemes from the existing lexicon, although the experimental results show improvements in the model's performance. In addition, data augmentation with various approaches is an excellent approach for increasing the number of validated datasets. In the future, we will try to increase the size of the validated dataset. An end-to-end approach could also be a potential solution for building a speech recognition model without constructing a lexicon. In addition, transfer learning using a pretrained model is also an interesting study to observe using Indonesian speech recognition datasets in the future.

Data Availability

The datasets used in this research are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was funded by the Institute for Research and Community Services, Universitas Syiah Kuala, under the Professor Research Grant 268/UN11/SPK/PNBP/2020. The authors also thank all validators who have helped them validate the speech datasets.

References

[1] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

[2] C. Hori and S. Furui, "A new approach to automatic speech summarization," *IEEE Transactions on Multimedia*, vol. 5, no. 3, pp. 368–378, 2003.

[3] T. Jo, "Text indexing," in *Text Mining. Studies in Big Data*-Springer, Berlin, Germany, 2019.

[4] K. A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, "Joint application of speech and speaker recognition for automation and security in smart home," in *Proceedings of the INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, 2011.

[5] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," 2015, https://arxiv.org/abs/1507.06947.

[7] W. Williams, N. Prasad, D. Mrva, T. Ash, and T. Robinson, "Scaling recurrent neural network language models," 2015, https://arxiv.org/abs/1502.00512.

[8] D. Hoesen, C. H. Satriawan, D. P. Lestari, and M. L. Khodra, "Towards robust Indonesian speech recognition with spontaneous-speech adapted acoustic models," *Procedia Computer Science*, vol. 81, pp. 167–173, 2016.

[9] A. Winursito, R. Hidayat, and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," in *Proceedings of the 2018 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 2018.

[10] U. N. Wisesty and W. Astuti, "Feature extraction analysis on Indonesian speech recognition system," in *Proceedings of the 2015 3rd International Conference on Information and Communication Technology (ICoICT)*, Nusa Dua, Bali, Indonesia, 2015.

[11] M. T. Uliniansyah, H. M. Riza, A. Santosa, M. Gunawan, and E. Nurfadhilah, "Development of text and speech corpus for an Indonesian speech-to-speech translation system," in *Proceedings of the 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, Seoul, Korea (South), 2017.

[12] S. Sakti, E. Kelana, R. Hammam, S. Sakai, K. Markov, and S. Nakamura, "Development of indonesian large vocabulary continuous speech recognition system within a-star project," in *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, Indonesia, 2008.

[13] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, "JTubeSpeech: corpus of Japanese speech collected from YouTube for speech recognition and speaker verification," 2021, https://arxiv.org/abs/2112.09323.

[14] Y. Long, Y. Li, Q. Zhang, S. Wei, H. Ye, and J. Yang, "Acoustic data augmentation for Mandarin-English code-switching speech recognition," *Applied Acoustics*, vol. 161, Article ID 107175, 2019.

[15] X. Sun, Q. Yang, S. Liu, and X. Yuan, "Improving low-resource speech recognition based on improved NN-HMM structures," *IEEE Access*, vol. 8, pp. 73005–73014, 2020.

[16] B. Oyo and B. Kalema, "A preliminary speech learning tool for improvement of African English accents," in *Proceedings of the 2014 International Conference on Education Technologies and Computers, ICETC*, Lodz, Poland, 2014.

[17] M. Qasim, S. Nawaz, S. Hussain, and T. Habib, "Urdu speech recognition system for district names of Pakistan: development, challenges and solutions," in *Proceedings of the 2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques, O-COCOSDA*, Bali, Indonesia, 2016.

[18] M.-G. Di Benedetto, J. Y. Choi, S. Shattuck-Hufnagel et al., "Speech recognition of spoken Italian based on detection of landmarks and other acoustic cues to distinctive features," *Journal of the Acoustical Society of America*, vol. 148, no. 4, p. 2808, 2020.

[19] V. M. Shetty, M. Sagaya Mary N J, and S. Umesh, "Improving the performance of transformer based low resource speech recognition for Indian languages," in *Proceedings of the ICASSP, IEEE International Conference on Acoustics*, Barcelona, Spain, 2020.

[20] S. Chu, S. Narayanan, and C. C. J. Kuo, "Environmental sound recognition using MP-based features," in *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA, 2008.

[21] K. M. Ravikumar, B. Reddy, R. Rajagopal, and H. C. Nagaraj, "Automatic detection of syllable repetition in read speech for objective assessment of stuttered disfluencies," in *Proceedings of World Academy of Science, Engineering and Technology*, vol. 22, pp. 270–273, 2008.

[22] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," 2016, https://www.danielpovey.com/files/2016_interspeech_raw.pdf.

[23] R. Serizel and D. Giuliani, "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children," *Natural Language Engineering*, vol. 23, no. 3, pp. 325–350, 2017.

[24] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Singapore, 2014.

[25] T. Zia and U. Zahid, "Long short-term memory recurrent neural network architectures for Urdu acoustic modeling," *International Journal of Speech Technology*, vol. 22, no. 1, pp. 21–30, 2019.

[26] M. D. Hassan, A. N. Nasret, M. R. Baker, and Z. S. Mahmood, "Enhancement automatic speech recognition by deep neural networks," *Periodicals of Engineering and Natural Sciences*, vol. 9, no. 4, pp. 921–927, 2021.

[27] B. Dendani, H. Bahi, and T. Sari, "Self-supervised speech enhancement for Arabic Speech recognition in real-world environments," *Traitement du Signal*, vol. 38, no. 2, pp. 349–358, 2021.

[28] V. Bhardwaj and V. Kukreja, "Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions," *Applied Acoustics*, vol. 177, Article ID 107918, 2021.

[29] J. Gnanamanickam, Y. Natarajan, and K. R. Sri Preethaa, "A hybrid speech enhancement algorithm for voice assistance application," *Sensors*, vol. 21, no. 21, p. 7025, 2021.

[30] A. Pandey, C. Liu, Y. Wang, and Y. Saraf, "Dual application of speech enhancement for automatic speech recognition," 2021, https://arxiv.org/abs/2011.03840.

[31] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," 2015, https://www.isca-speech.org/archive/interspeech_2015/peddinti15b_interspeech.html.

[32] V. Peddinti, G. Chen, D. Povey, and S. Khudanpur, "Reverberation robust acoustic modeling using i-vectors with

time delay neural networks," 2015, https://www.danielpovey.com/files/2015_interspeech_aspire.pdf.

[33] K. J. Han, J. Pan, V. K. N. Tadala, T. Ma, and D. Povey, "Multistream CNN for Robust Acoustic Modeling," 2021, https://arxiv.org/abs/2005.10470.

[34] M. Kubanek, J. Bobulski, and J. Kulawik, "A method of speech coding for speech recognition using a convolutional neural network," *Symmetry (Basel)*, vol. 11, no. 9, pp. 1185–1212, 2019.

[35] I. S. Areni and A. Bustamin, "Improvement in speech to text for bahasa Indonesia through homophone impairment training," *Journal of Computers*, vol. 28, no. 5, pp. 1–10, 2017.

[36] K. Nugroho, E. Noersasongko, M. Purwanto, D. R. I. M. Setiadi, and D. R. I. M. Setiadi, "Enhanced Indonesian ethnic speaker recognition using data augmentation deep neural network," *Journal of King Saud University—Computer and Information Sciences*, vol. 34, no. 7, pp. 4375–4384, 2022.

[37] D. Povey, "The kaldi speech recognition toolkit," 2011, https://infoscience.epfl.ch/record/192584.

[38] T. F. Abidin, A. Misbullah, R. Ferdhiana, M. Z. Aksana, and L. Farsiah, "Deep neural network for automatic speech recognition from Indonesian audio using several lexicon types," in *Proceedings of 2020 International Conference on Electrical Engineering and Informatics (ICELTICs)*, Aceh, Indonesia, 2020.

[39] D. Povey, G. Cheng, Y. Wang et al., "Semi-orthogonal low-rank matrix factorization for deep neural networks," 2018, https://www.isca-speech.org/archive/interspeech_2018/povey18_interspeech.html.