IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 3, 2021, accepted April 21, 2021, date of publication May 4, 2021, date of current version May 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077441

Improved Transcription and Speaker Identification System for Concurrent Speech in Bahasa Indonesia Using Recurrent Neural Network

MUHAMMAD BAGUS ANDRA, (Member, IEEE), AND TSUYOSHI USAGAWA, (Member, IEEE)
Department of Computer Science and Electrical Engineering, Kumamoto 860-0862, Japan

Corresponding author: Muhammad Bagus Andra (andra@ieee.org)

This work was supported in part by the JSPS KAKENHI Grant-in-Aid for Scientific Research under Grant 19H1225100.

ABSTRACT Bahasa Indonesia is one of the most prominent low-resource Languages that still lack development in regards to communication-assisting technology. This paper proposes an improved system for generating transcript and identifying speakers from a concurrent speech in Bahasa Indonesia. The proposed method is applicable in a situation such as an online meeting and remote conference. The system combines Reinforced Learning (RL) Model with pitch-aware speech separation to identify the speakers in a concurrent speech. A Recurrent Neural Network (RNN) is utilized to generate the text transcript which is later improved by an external language model and spelling correction model. The proposed system was able to identify up to 5 speakers with a variable degree of confidence and generate a transcript for each of them with better quality compared to other methods when evaluated with several metrics. The result shows that the proposed method perform better compared to the baseline method, even in the single-speaker situation, and function in the simultaneous-speech situation, with an average Word Error Rate (WER) of 16.59% for two speakers, 26.72% for three speakers, and 31.50% for four speakers.

INDEX TERMS Bahasa Indonesia, deep learning, Recurrent Neural Network, speech processing, speech separation.

I. CHARACTERISTICS OF INDONESIAN LANGUAGE

Indonesian Language, or is often called Bahasa Indonesia is a unity language that belongs to Austronesian family formed from hundreds of local languages throughout the country. While it is formed from a wide variety of ethnic accents, often words share a similar pattern and meaning across many places. Compared to other languages which have high percentage of the native speaker, Bahasa Indonesia is spoken as mother tongue only by 7% of it's population. And more than 158 million people speaks it as the secondary language with various proficiency [1]. There are estimated 300 ethnic groups living in more than 17,000 islands, speaking 365 native languages and no less than 669 dialects [2].

Modern Indonesian Language is derived from the Malay which was the main language used in the southeast Asia. It is closely related to the Malay spoken in Malaysia, Singapore and Brunei but it adopted different orthography. It is appointed as the national language after the declaration of independence in 1945 which become the standard language that can be spoken in every part of Indonesia [3].

Compared to other languages, such as Chinese, Bahasa Indonesia is not tonal language which means the difference in pronunciation tone and pitch has no effect to its meaning. Compared to European languages, Indonesian really few usage of gendered words. The verbs also don't take different form for showing number, person or tense. For expressing plural, Bahasa Indonesia use the means of repetition of word. It is also considered as a member of agglutinative language family, meaning that it has wide range of prefixes and suffixes. [3]

According to [4] Bahasa Indonesia has 33 phenomes which consist of seven vowel phonemes, three diphthong and 23 consonant phonemes. These phonemes are the standard phonemes used by Indonesians when uttering Indonesian words without considering their allophone. Table 1 shows the list of all phonemes used in Bahasa Indonesia.

The associate editor coordinating the review of this manuscript and approving it for publication was Emre Koyuncu.

TABLE 1. List of phonemes used in Bahasa Indonesia.

No.	Phonetic Category	Phoneme
1.	Vowel	/a/
2.		/i/
3.		/u/
4.		/é/
5.		/ê/
6.		/è/
7.		/o/
8.	Diphtongs	/ai/
9.		/au/
10.		/ou/
11.	Consonant	/b/
12.		/P/
13,		/T/
14,		/d/
15,		/g/
16.		/h/
17.		/j/
18.		/k/
19.		/m/
20.		/l/
21,		/N/
22.		/c/
23.		/R/
24.		/S/
25.		/W/
26.		/Y/
27.		/Z/
28.		/q/
29.		/Ng/
30.		/Ny/
31.		/Sy/
32.		/x/
33.		/f/

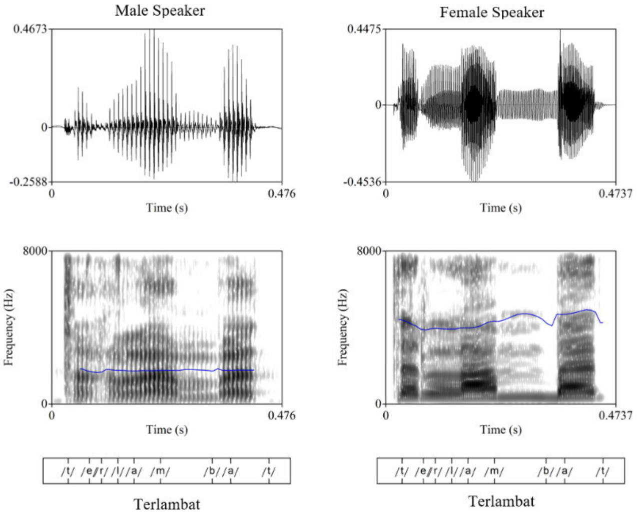

FIGURE 1. Example of Bahasa Indonesia speech signal from male and female speaker uttering word "terlambat" (late) with its respective pitch contour (shown by blue line in the spectrogram) and phoneme sequence.

Vocabulary in Bahasa Indonesia receive a lot of influence from foreign cultures that have passed through the land from the history. Many of the words bear the resemblance to its root counterparts from Indian, Chinese, Arabic, Portuguese, Dutch and English. Modern Indonesia is written in Roman script that consist of 26 letters from 'a' to 'z'.

It has a highly phonemic orthography, meaning that almost all graphemes represent one phoneme sound, except for a few sounds represented by diagraphs and vice versa, almost all phonemes are represented by either one or two graphemes. Some examples of uttered speech in Bahasa Indonesia with its respective phoneme and grapheme label are shown in Figure 1.

II. SPEECH PROCESSING IN INDONESIAN LANGUAGE

In this we section we explore several researches in the field of speech processing in Indonesian Language and the approach that has been proposed to tackle the challenge in developing speech technology for low-resource language which also found in Indonesian Language.

A. PREVIOUS RESEARCH

A great deal of past research in Bahasa Speech focused on automatic speech recognition with a traditional framework. For example, Muljono *et al.* propose using Sphinx4 for continuous speech recognition. Their research includes the combination of the Indonesian phoneme and acoustic model to be used in Sphinx4 [5]. However, the testing data that was used consists of only greeting sentences and has not been verified in more general cases. In contrast, C. H. Satriawan *et al.* suggest using the mel-frequency cepstral coefficients (MFCCs) and predictive linear prediction (PLP) feature of the speech [6], while B. D. Trisedya established a Graph Annotation Format (GrAF)-compliant Indonesian speech recognition web service [7]. The GrAF, one of the formats that implement the conceptual standard annotation of the Language Annotation Framework (LAF), applies graph theory to model the linguistic annotation that can facilitate creating and representing several annotations and incorporating them into a single and integrated annotation. Both of these methods also use Sphinx4 for the Automatic Speech Recognition (ASR) framework.

Some modern approach has also been considered by researchers, such as research by Rifqi Adiwidjaja and M. I. Fanany which introduced end-to-end Indonesian speech recognition, which uses the convolutional layer and gated recurrent units as the hidden layer [8] In their proposed method, the ResNet layer extracts the spatial feature of the speech from the spectrogram and passes it to the Bi-GRU layer. The prediction is then carried via the connectionist temporal classification function. Another method proposed by B.T Atmaja using Time-Delay Neural Network (TDNN) which were trained on MFCC feature extracted from Bahasa Indonesia Corpus. The network exploits the TDNN architecture which transform the feature on narrow context on early layers and wider temporal context on deeper layer [9]. Some research that also utilizes deep learning architecture focuses more on a specialized topic, such as that of Atmaja and Akagi [10] and Lasiman and Lestari [11], which concerns recognizing emotion in speech; Citta Anindya *et al.*, which concerns recognizing speech as robot command [12]; and E. R. Swedia, which specifically concerns recognizing digits [13].

B. CHALLENGE IN SPEECH TECHNOLOGY DEVELOPMENT

The development of assistive communication and speech technology for low-resource languages are still hindered by many problems. Factors such as general scarcity of the corpus, insufficient variety of the recorded speech, and low quality of the recording contribute a lot to slower research and development of such a system for these languages.

These problems are also found in case of Bahasa Indonesia, compared to other major language, it is seldom explored and researched which results in slower development of technology like automatic speech recognition (ASR), text-to-speech, and automatic transcription systems, which could be useful for discussion reviewing and archiving purposes.

One of the most influential corpus for Bahasa Indonesia was developed by Sakriani Sakti et. al. as a part of Asian Speech Translation (A-STAR) project [3]. This corpus contains recorded speech from selected newspaper article and telephone application resulting in total of 5668 phonetically balanced unique sentences (79.5 hours of speech). Another noteworthy corpus is Tokyo Institute of Technology Multilingual Speech Corpus-Indonesian (TITML-IDN) produced by Koichi Shinoda and Sadaoki Furui [14], composed by 343 unique phonetically balanced sentences (around 20 hours of speech) and Under-Resourced Bahasa Indonesia Speech Corpus by Elok Cahyaningtyas and Dhany Arifianto [15]. This corpus was constructed from 1029 declarative sentences and 500 question sentences gathered from movie and drama transcript. It was spoken by 6 professional announcers resulting in total of 10 hours of speech.

When one compares the content of the Indonesian Language corpus to the more widely-known such as switchboard [16] or LibriSpeech [17] it is become apparent that there is large gap in data and information contained. Even when all the above Indonesian Language Corpus is combined it only contains about 42% speech volume of switchboard corpus (109.5 hours vs 260 hours of speech). Moreover, the labelling convention and recording environment differ largely between corpus which makes it harder to process and standardize.

To overcome this hurdle, many researchers resort to deep learning approach that allows flexible model definition and advanced learning technique that fits the characteristics of the target language. This applies for many low-resource languages that face similar obstacle. For example, the end-to-end speech recognition model for Tibetan that was developed by Xiaojun and Heming [18] takes the advantage of knowledge transfer to use the model that is originally trained with Chinese and English phoneme to recognize the speech from Tibetan corpus. A similar approach is also proposed by Weizhao et. al. in their Mandarin-Tibetan speech synthesis model [19], which uses Long Short-term Memory (LSTM) variant of Recurrent Neural Network (RNN) instead of Listen, Attend and Spell (LAS) model in the former research as it is more suitable for training voice model and generating acoustic parameter.

A promising result has also been shown by Danyang et. al [20]. Jiangyan Yi et. al. [21], and Chien-Ting Lin [22] models that utilize a similar combination of Deep Neural Network (DNN) with shared hidden layer for speech recognition in various languages such as Vietnamese, Turkish and Mandarin.

Another approach to compensate for the lack of required corpus is by the means of data augmentation which is shown to have positive result in the research by P.N. Hadiwinoto and D. P. Lestari that incorporates dictation speech as parallel corpus to augment the data in the spontaneous speech corpus.

III. PROPOSED METHOD OVERVIEW

While the majority of past research addresses numerous speech aspects, it relies on classic methods and frameworks, such as Sphinx4 and HMM based model, which could still be improved in term of its performance and accuracy. Meanwhile, more modern approach has limited use for a specific case and condition. Moreover, many of these studies consider only a single-speaker condition, which may not function in concurrent speech situations, including online discussions. Considering numerous issues from the previous method the objective of this research is to propose a system that:

1) Could take the advantage of the cutting-edge performance of Neural Network architecture while keeping its generality.
2) Perform well albeit trained with limited amount of data
3) Works as end-to-end system that could eliminate segmentation or alignment problem as well as word with spelling variants.
4) Able to handle separate speech in a concurrent speech situation and generate transcript for each individual speech

To achieve this, we employed a several methods in our system. First, the extended Kalman filter is employed for enhancing the speech by reducing noise from the recording environment. A pitch-aware speech separation system and speaker identifier are incorporated to enable the system to work in a concurrent speech situation while a deep recurrent neural network (deep RNN) was used to generate a transcription of each separated speech signal. Post-processing of the transcript was completed with the help of a dictionary, WordNet Bahasa Indonesia [23], and a spelling correction model. Figure 2 depicts the overall transcription and speaker identification system that is presented in this paper.

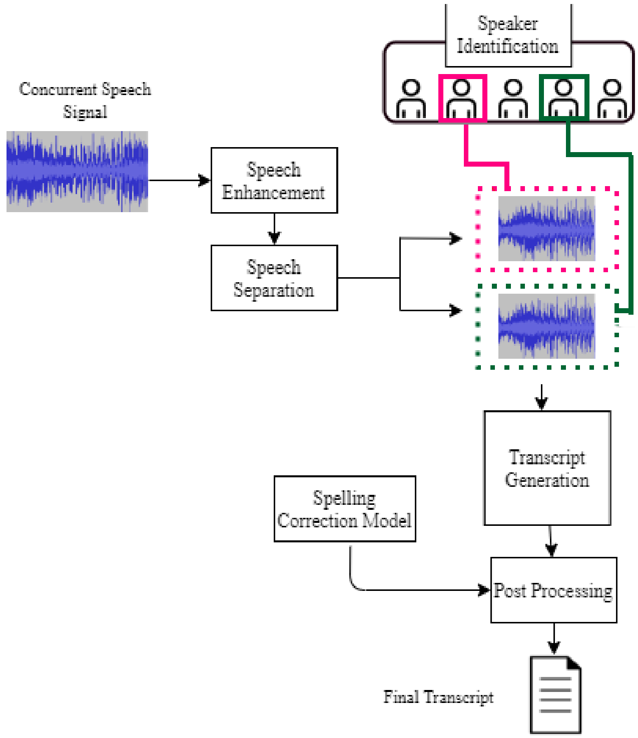

FIGURE 2. Proposed method diagram. The input mixed speech signal undergoes four main processes or stages for generating a transcript.

IV. SPEECH ENHANCEMENT WITH EXTENDED KALMAN FILTER

The intelligibility of the speech is key feature for speech recognition system to be able to distinguish difference in words. However, in the real-life situation, the speech quality is often degraded by various interference such noise from environment and recording device. To preserve the quality of the speech recording, we employed speech enhancement as a pre-processing step in the system. This process is done for both speech data used in training step and prediction step.

In order to achieve this, one could resort to a well-known technique such as Linear Predictive Coding (LPC) based enhancement algorithm. This technique relies on the available clean speech to accurately estimate the LPC. However, it is not suitable for our application where multiple persons are speaking in different environment with unpredictable noise nature [24]. This limitation also applies to spectral subtraction-based technique which require the spectral pattern of the noise to recover the clean speech. While some improvement has been made for this technique that uses spectral estimator for adaptive noise reduction [25], [26], the performance of this technique is still heavily depends on the estimation of noise spectrum which is difficult to do in low SNR condition.

In order to overcome such limitation, we resort to iterative variation of extended Kalman filter for speech enhancement. In spite of its initial application in spacecraft and aircraft signal analysis [27], Kalman filter has been actively researched for speech enhancement. Compared to other methods it has advantage such as the ability to maintain non-stationary nature of the speech and it does not need to assume stationary condition within small analysis window [28]. Unlike to the traditional LPC based enhancement algorithm. Iterative Kalman filter uses sequential estimation technique for estimating LPC and noise variance from noisy speech which is suitable for our application where clean speech is not available and the noise feature is unknown beforehand.

As per Sharon Gannot [29], to apply the Kalman filter to the speech signal, one must treat the clean speech signal as an autoregressive process that can be described as an all-pole FIR filter, as this equation illustrates:

$$x\,(k) = \sum_{i=1}^{p} a_i x\,(k-1) + u\,(k) \tag{1}$$

and the noisy speech is defined as:

$$y\,(k) = x\,(k) + v(k) \tag{2}$$

where $x\,(k)$ is the k^{th} sample of the clean speech, $y\,(k)$ is the k^{th} sample of the noisy speech and a_i is the i^{th} coefficient of LPC coefficient $u\,(k)$ and $v\,(k)$ is unrelated process noise. This system can then be represented in state-space model, where Φ, $\boldsymbol{H}$, and $\boldsymbol{G}$ represent vectors or matrices. The state equation is given below:

$$s\,(k) = \Phi s\,(k-1) + \boldsymbol{G}u(k) \tag{3}$$

Observation equation:

$$y\,(k) = \boldsymbol{H}s\,(k) + v(k) \tag{4}$$

Given the above equation, the Kalman filter estimates the state vector of $\overset{\frown}{\hat{s}}\ (k|k)$ from the corrupted speech with these equations, first the initialization of the state-space vector:

$$\hat{s}(0|0) = 0 \tag{5}$$

$$\sum\nolimits_s (0|0) = [0]_{pxp} \tag{6}$$

Time update (predictor):

$$\hat{s}\,(k|k-1) = \Phi\hat{s}\,(k-1|k-1) \tag{7}$$

$$\sum\nolimits_s (k|k-1) = \Phi \sum\nolimits_s (k-1|k-1)\,\Phi^{\boldsymbol{T}-} + \boldsymbol{G}\sigma_u^2\boldsymbol{G}^{\boldsymbol{T}} \tag{8}$$

Measurement update (corrector):

$$e\,(k) = y\,(k) - \boldsymbol{H}\hat{s}\,(k|k-1) \tag{9}$$

$$\boldsymbol{K}\,(k) = \sum\nolimits_s (k|k)\,\boldsymbol{H}^{\boldsymbol{T}-}(\boldsymbol{H}\sum\nolimits_s (k-1|k-1)\,\boldsymbol{H}^{\boldsymbol{T}} + \sigma_u^2)^{-1} \tag{10}$$

$$\hat{s}\,(k|k) = \hat{s}\,(k|k-1) + \boldsymbol{K}\,(k)\,e(k) \tag{11}$$

$$\sum\nolimits_s (k|k-1) = (\boldsymbol{I} - \boldsymbol{K}(k)\boldsymbol{H})\sum\nolimits_s (k|k-1) \tag{12}$$

Finally, the estimated enhanced speech (at time k):

$$\hat{x}\,(k) = \boldsymbol{H}\hat{s}\,(k|k) \tag{13}$$

The above procedure are repeated for the every speech frames and continued until all frame is processed. At the end of the processing the final enhanced speech $\hat{x}\,(k)$ is then obtained [28]. As for the iterative Kalman filter, the process is also done frame by frame, however, it contains two loops of iteration. In the inner loop, the state space model parameters of the Kalman filter are updated sample-by-sample. The interfering noise components are reduced significantly when the inner loop is completed for the entire frame. The outer loop iterative process stops when the filter converges or the preset maximum number of iterations is finished. To illustrate the effect of the Kalman filter, Figure 3 shows the example of spectrogram for the clean, noisy and enhanced speech.

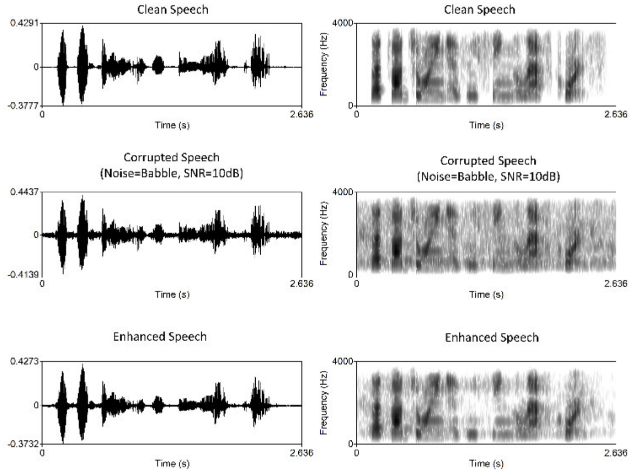

FIGURE 3. Speech signal in time domain and its respective spectrogram The first plot shows the clean speech, the second plot shows a degraded speech and the last plot shows the speech after enhancement process.

As could be seen in the plot comparison above, the iterative Kalman filter method was able to eliminate considerable amount of environment noise, which in this case babble noise. and reconstruct the estimation of clean speech which improve the intelligibility of the spoken word.

To further evaluate the effect of the Kalman filter, we use NOIZEUS speech corpus dataset [30]. The dataset contains clean and noisy speech that is degraded with various type of noise. To simulate the situation of online discussion participant that joins from various environment we decide to choose babble, street, restaurant and airport noise as the interference. We took 10 speech utterance and use the range of 0dB to 15dB SNR for each of the noise type. We apply the Kalman filter to the chosen speech and calculate the Segmental SNR (SegSNR) value of the enhanced speech using the clean speech as the baseline. We compare the performance of the Iterative Kalman filter with traditional LPC, spectral subtraction, and Wiener filtering. Figure 4 shows the performance comparison between each method.

From the result we could see that each method perform differently with depends on the type of the noise. For example, spectral subtraction method performs really well in the case of airport noise interference and in many case Wiener filter perform the worst with low input SNR but getting better when the input SNR is high. However, in the case of iterative Kalman filter, the performance is constant across various input SNR and noise types. It also outperforms all of the other methods in every scenario.

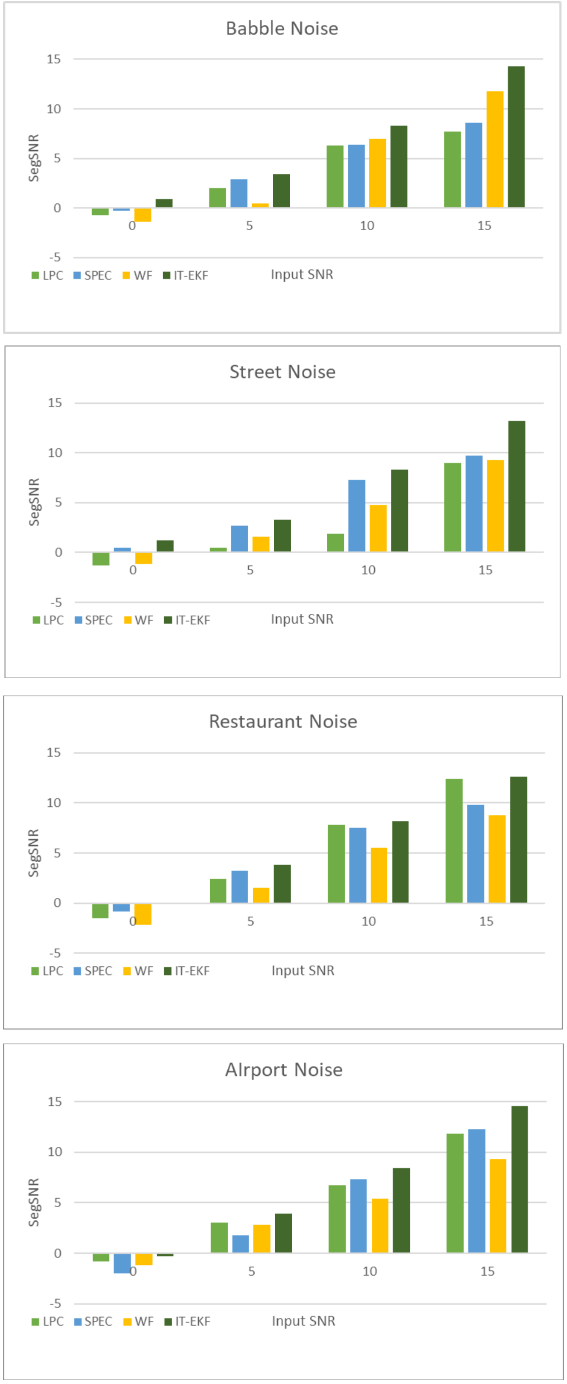

FIGURE 4. Comparison of speech enhancement method in a degraded speech exposed to four types of different noise each with various input SNR. The iterative Kalman filter perform the best in every scenario.

V. SPEECH SEPARATION AND SPEAKER IDENTIFICATION

For the system to generate transcript for individual speaker, speech separation is crucial step in the system. For our application, the input speech will be mainly single channel mixed speech with a possibility of more than one speaker speaking at the same time (concurrent speech). Single speech separation is much harder to do when compared to binaural or multi-channel speech separation as we cannot exploit the time-difference or phase-difference feature found in those situation [31]. Ideally the system should also be able to identify each of the speaker so that it can match the generated transcript to each of the speaker, In this section we describe in detail the process of speech separation and speaker identification.

A. PITCH AWARE SPEECH SEPARATION

With the rapid development of complex deep learning architecture, the area of single speech has seen a substantial progress throughout the year. Several methods such as deep clustering and permutation invariant training has been proposed with a great result [32], [33]. A large-scale network that combines both methods such as Computational Auditory Scene Analysis (CASA) [34] and chimera network [35] has been proven to achieve state-of-the-art performance. However, as described by [36], the current proposed method still has a problem when separating combined speech from same gender. Especially, female-female (FF) combined speech has been shown to have worse performance than opposite gender mixture (MF).

In the situation like online meeting or discussion where any mixture of speech could happen, it is important to consider maximizing performance for any possible speech combination. One consideration is to incorporate pitch information by the means of pitch tracking. This decision is also supported by the fact that Indonesian Language is not a tonal language as described in section I, meaning that the variation in the pitch could be exploited for separation feature while the meaning of the utterance is unaffected. Figure 5 shows the pitch contour by three male speakers speaking same word formed from fundamental frequency (F0).

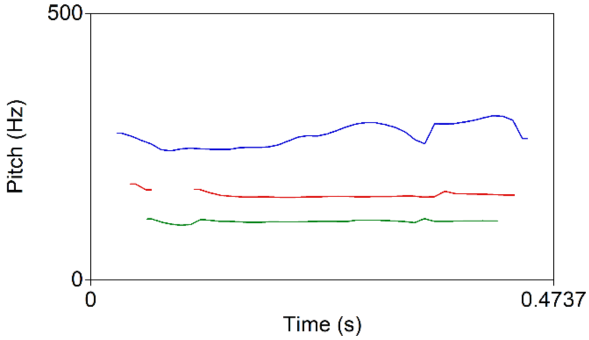

FIGURE 5. Example of pitch contour for same word, uttered by three different male speakers.

The separation process is done as described by Wang, Soong, and Xie [36]. In the first stage of the separation, a Neural Network based deep clustering model [32] is trained to do deep embedding for speech feature. The model is then trained to do clustering by learning a mask for each source, this clustering process will group the speech feature based on their mask. After masking, another component of the model will perform pitch tracking for each speech source. Finally, in the second stage of the separation, the trained model will be augmented with the corresponding combination as the input for the final separation stage. The Overall process of the speech separation is illustrated by the diagram in Figure 6.

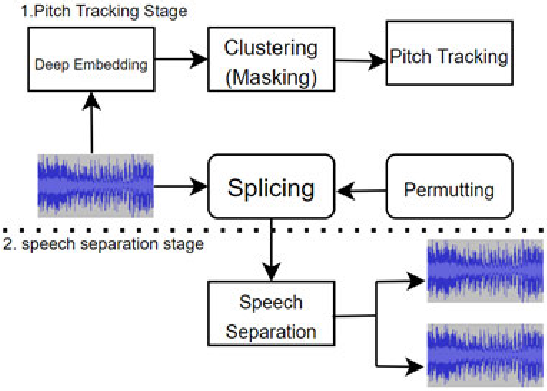

FIGURE 6. Diagram of speech separation process.

Following deep clustering model [32], the architecture for speech separation uses 4 layers of bidirectional LSTM unit (BLSTM) followed by a 2 fully connected feed-forward layers which forward the result to the output layer. The input feature uses a 129-dimensional vector calculated from 256-point STFT. The parameter used for the network is listed in Table 2 and the architecture is illustrated by Figure 7.

TABLE 2. Parameter list for speech separation architecture.

Parameter	Value
# BLSTM Layer	4
# MLP-Feed forward	2
# BLSTM unit	896
# MLP unit	300
Learning rate	2×10^{-7}
Activation function	ReLU
Batch size	100
Dropout rate LSTM (#1-#3)	0.5
Dropout rate LSTM (#4)	0.3
# Epoch	150

For the training process, WSJ0-2mix corpus [37] is used, the corpus contains 20,000 two-speaker mixtures for the training set. The test set is used for evaluation with disjointed speaker set from the training set. The training was done for 150 epochs and the progress of Mean Square Error (MSE) is observed throughout the process. The graph illustrated by Figure 8 shows the progress of the model training.

Evaluation process is done by measuring the Signal to Distortion Ratio (SDR) which is a scale-invariant SNR that indicates the improvement of the speech signal.

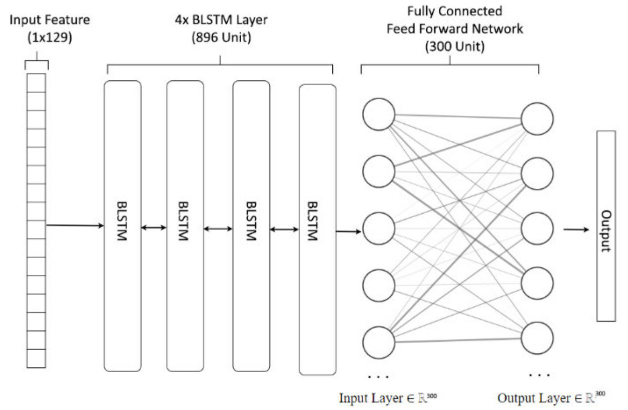

FIGURE 7. Architecture for pitch-aware separation network based on deep clustering model which employs layers of LSTM that is connected to feed-forward network.

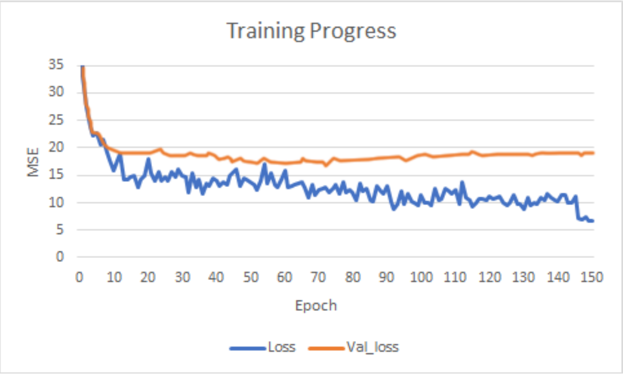

FIGURE 8. Training progress for speech separation model, showing the change of MSE over epoch.

TABLE 3. Speech separation performance for various mixture.

Method	Mixture		
	FF	MM	FM
DPCL	11.8	12.2	12.9
DPCL-PIT	13.2	13.5	13.8

For the evaluation, a permutation of mixed speech of same gender (M-M), (F-F) and opposite gender (M-F) are considered. We compare the performance between original deep clustering model (DPCL) with the one augmented with pitch tracking process (DPCL-PIT). It is all done under the *oracle* scheme where the procedure of mixture is known and the pitch of the source uses highest SNR. The result is shown in Table 3.

Our evaluation shows that incorporating pitch information to the model increase its performance in term of SDR. The significance could be seen especially for the same gender mixture where it increased by 1.4 for female-female mixture. And 1.3 for male-male mixture. The opposite gender mixture also sees an improvement of 0.9 SDR compared to the base DPCL model.

B. SPEAKER IDENTIFICATION

In addition to speech separation, the system is also required to identify each speaker identity. While this process does not directly contribute to the system's accuracy in generating transcript, the ability to assign the generated transcript to each individual will increase the practicality of the system in its application for archiving the dialogue and could save a lot of time compared to manual assignment of the transcript.

The speaker identification, much like pitch tracking process described in the previous section is a speaker adaptation process that uses a specific feature from the speech signal to differentiate one individual from others. This adaptation process usually involves collecting some sample speech from each individual speaker.

The identification model used in this research is adapted from interactive speech recognition model described in [38]. This model consists of two parts, the guesser and the enquirer which are tasked to maximize the probability of guessing the speaker identity from the given list of speakers. This problem is formally defined as follows: K available guests are characterized by their voice print $\mathrm{g} = \left[\mathrm{g}^{\mathrm{k}}\right]_{\mathrm{k}-1}^{\mathrm{K}}$ the enquirer seeks to build a list of words $\boldsymbol{w} = [w_t]_{t-1}^{t}$ for the guesser to maximize the probability of $x \in g^k$. By repeating this process in the manner of Reinforcement Learning (RL), the model could increase its accuracy and confidence in guessing the speaker incrementally. However, unlike the original model, the enquirer in the proposed model didn't ask for the guest to speak a certain word, instead, it picks the words from the guest's continuous speech that can maximize the function automatically. Figure 9 and 10 illustrate the guesser and enquirer model while the detailed parameter used for each model is listed in Table 4 and 5.

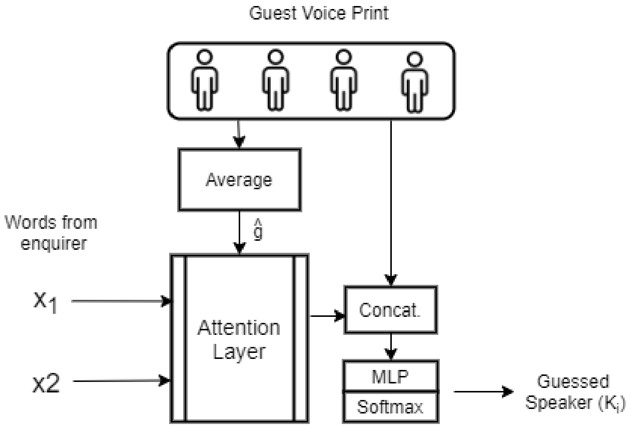

FIGURE 9. Guesser component uses words fed from enquirer and guest voice print to identify the speaker.

For the training process of the model. TIMIT corpus was used [39]. Mel Frequency Cepstral Coefficient (MFCC) is extracted from the khz downsampled speech and processed through X-vector network to obtain a 128-dimensional feature vector. For the guesser model, ADAM optimizer is used to minimize the cross-entropy over 400 epochs.

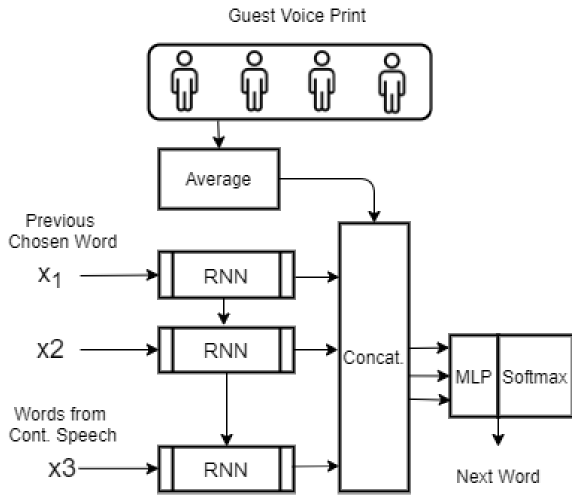

FIGURE 10. Enquirer component takes the utterance of the guest and combines it with previous words for to generate next words for guesser.

TABLE 4. Parameter list for guesser model.

Parameter	Value
# Attention Layer	1
# MLP-Feed Forward	1
# Attention Layer unit	256
# MLP unit	512
Learning rate	3×10^{-4}
Activation function	ReLU
Batch size	1024
# Epoch	400

TABLE 5. Parameter list for enquirer model.

Parameter	Value
# RNN	3
# MLP-Feed Forward	1
# Attention Layer unit	128
# MLP unit	256
Learning rate	5×10^{-3}
Activation function	ReLU
Batch size	512
# Epoch	750

The enquirer model uses PPO with ADAM optimizer to maximize the reward encoded as the guesser success ratio rate over 750 epochs. The progress of the training is shown by Figure 11 and 12.

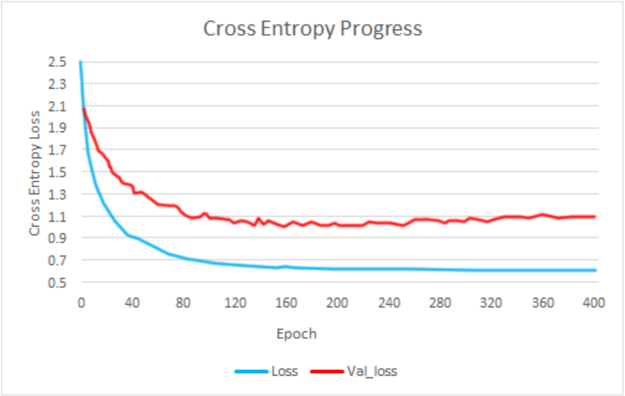

FIGURE 11. Training progress for guesser model, showing the change of cross entropy loss over epoch.

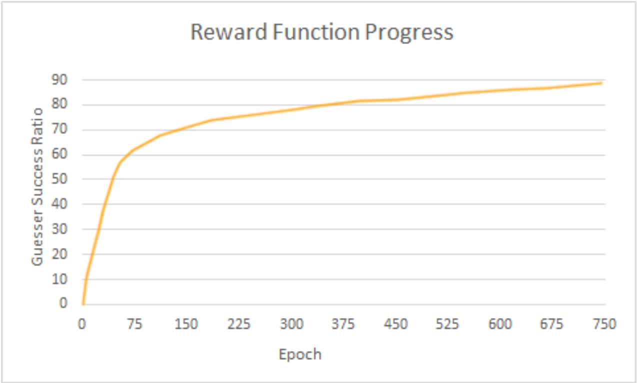

FIGURE 12. Training progress for enquirer model, showing the change of reward function over epoch.

VI. TRANSCRIPTION GENERATION

In this step, the system will generate word transcript for each individual speech signal obtained from the separation process. Generally, traditional approach uses a acoustic model to transform feature of a speech into probabilistic model which is used to predict likelihood estimation of n-gram. Finally, language model are used to align the output to a phonemic sequence, however as we have explained in the section 1, this approach has a severe limitation in predicting similar utterance and prone to misalignment due to the existence of segmentation issue [40].

Recently, machine learning approach has been widely explored in speech recognition area that is capable to solve these issues. While the general idea behind the procedure is the same, machine learning models are capable of doing complex mapping between speech input and desired output, whether it is direct word sequence or phoneme models. Many models also allow the network to learn contextual information in the speech, enable the model to do recognition for multiple language and accurately recognize words with spelling variant without incorporating additional language model.

Our system uses sequence-to-sequence framework which is part of end-to-end approach where the input speech is directly mapped to a word sequence. Specifically, it is based on Listen, Attend and Spell model (LAS) proposed by William Chan et. al. [41] that consist of two parts, the encoder and predictor (or decoder) both of which are interconnected RNNs. The encoder works to transform the speech signal into a high-level vector representation, much like an acoustic model while the decoder part of the network transforms the high-level vector by conditioning on the previous target to estimate the label for the output layer. Figure 13 shows the general architecture of the transcript generation module.

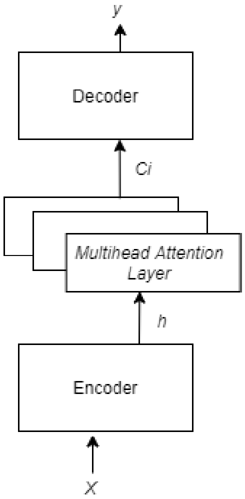

FIGURE 13. Transcript generation architecture based on LAS model it consists of 3 parts interconnected to each other.

As it has been mentioned before, the encoder part of the model takes the acoustic input feature extracted from the speech signal which then mapped to higher-level representation. h. It is then processed through attention layer which determines which features in h should have been attended to estimate the next output sequence. In the last part, decoder takes the attention context, c to produce a probability distribution $P(y_i|y_{i-1}, \ldots, y_0, x)$, over the current sub-word unit, y_i, given the previous units $\{y_{i-1}, \ldots, y_0\}$, and input. x. The following subsection explains each component of this model in detail.

A. ENCODER

For the encoder part of this model. A BLSTM RNN is employed. The objective of this network is to reduce the dimensionality of the input feature by transforming it to a higher-level feature h. This process is essential in the application such as speech recognition where the input sequence (speech signal frames) is far longer than the output sequence (word).

This network uses the pyramidal structure which has the advantage of effectively extract the relevant information from the input feature which helps the model to converge faster, compared to the regular BLSTM [42]. The architecture of the encoder network is illustrated by Figure 14.

For this network, 3 hidden layers are used on top of the bottom BLSTM layer that have 512 unit. Compared to the previous model [43], Contextual layer trajectory LSTM is employed. For input feature, an improved PLP based feature [44] are calculated from the individual speech. This feature is interpolated with 5 frames to the left.

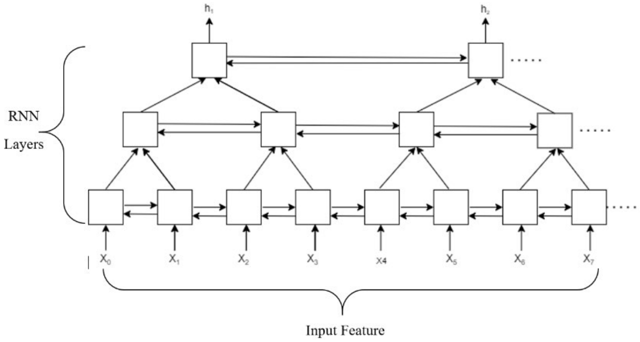

FIGURE 14. Encoder network that takes input speech and transform it into high-level feature, it employs a total of 4 BLSTM layer with pyramidal construct.

B. MULTI-HEAD ATTENTION LAYER

Multi-head Attention (MHA) is a mechanism where several attention heads are used to generate multiple attention distribution towards the input feature. This method has been utilized in other sequence-based model such as machine translation and has been shown to improve the decoder module to learn contextual information from the feature. This also helps the model learn multiple spelling variants that often found in number spelling such as "tiga puluh tujuh" which could mean a number 37 or a number 7 that repeats 30 times.

Compared to the single attention mechanism where the model relies on the encoder to pass the information about the utterance, multi-headed attention distributes this task to several attention head which each point out a different area of attention within the same feature that enables the decoder to have a multiple context to process. We employ additive attention for the proposed model which uses 4 attention head.

C. DECODER

The decoder receives high-level feature as an input and uses attention context from the attention layer to compute probability distribution of the next character based on all characters that have been seen in the previous step. For each time step the attention layer produces a context coefficient that contains the information of the speech signal required to predict the next character in a word sequence. The state of the decoder is then matched to the input feature to generate context vector and probability distribution over the current word. For the inference step, the objective of the decoder is to find the most likely character sequence that forms word as described by equation below:

$$\overbrace{y} = \arg\max_y \log P\left(y|x\right) \tag{14}$$

Decoding process is done with a beam search algorithm. At each iteration, every possible character is added to the beam based on the probability distribution until end of word token is encountered. However, only characters with most likely probability is kept and the rest are removed.

Compared to previous models [43], we choose grapheme sequence for the output instead of wordpiece and phoneme.

This decision is supported by the similar performance between each feature as shown by [45] and the nature of Bahasa Indonesia where each phoneme represent one or two graphemes. This connection enables the model to bypass mapping process between phoneme to grapheme and use grapheme directly as the output.

For the decoder network, a 2-layer unidirectional LSTM layer is used, each layer contains 1024 unit and softmax function is used for predicting the probability distribution. Figure 15. Illustrates the decoder network.

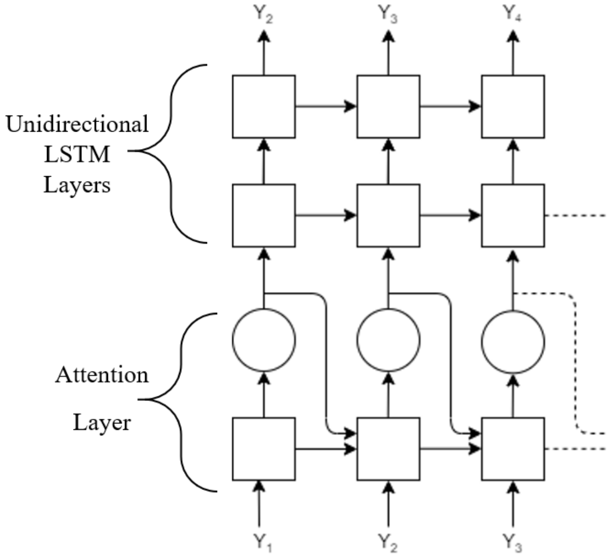

FIGURE 15. Decoder network generate the probability distribution for the grapheme character sequence using contextual information from the attention layer. It consists of 2 LSTM layers.

D. TRAINING

The training process for the transcript generation is done in joint, where the decoder and encoder part are trained synchronously. For the optimization, MWER loss [46] is adapted instead of grapheme sequence error or CTC loss often used in CTC based system. MWER loss is a sequence-level loss function that aims to minimize the number of word errors. The loss function is shown in equation below:

$$\mathcal{L}_{MWER} = E_{P(y|x)}\left[\mathcal{W}\left(y, y^*\right)\right] + \lambda\mathcal{L}_{CE} \tag{15}$$

where $\mathcal{W}(y, y^*)$ denotes the number of word errors in the sequence estimation y compared to the ground truth y^* which is interpolated by the standard cross-entropy based loss $\lambda\mathcal{L}_{CE}$.

Bahasa Indonesia speech dataset which consists of 29828 single channel speech utterance is used for training dataset. Speech enhancement method as described in Section IV is applied to the training dataset and a 80-dimensional modified PLP feature [44] is calculated which is used as the input feature. It uses 30ms window and interpolated with 5 frames to the left. The parameter used for the network is listed in Table 6.

TABLE 6. Parameter list for transcript generation model.

Parameter	Value
Encoder # BLSTM Layer	4
Encoder # BLSTM Unit	1024
Encoder Activation Function	Sigmoid
Decoder # BLSTM Layer	2
Decoder # BLSTM Unit	512
Encoder Activation Function	ReLU
Attention Layer	1
Attention Head	4
Learning rate	0.1 (0.85 decay/1000 word)
Activation function	ReLU
Batch size	1024
# Epoch	550

The decoder and encoder are trained for 550 epochs and MWER loss progress is observed. Figure 20 shows the training progress for each dataset.

VII. TRANSCRIPTION POST PROCESSING

Since the transcription process for the RNN is performed at sentence level and without a language guideline, some words are often misspelled. In some cases, the words are recognized as being similar to their pronunciations. Confusion can also result when a short pause occurs between spoken words so that two words are detected as a single word.

To minimize this kind of error, one can assign a score to the transcribed sentence with the language model and find the alternative sentence with the maximum score. WordNet Indonesia [23] is then used to estimate an Indonesian thesaurus, as the dataset used in this study are all in Indonesian.

The scoring function, made from WordNet itself, is calculated by obtaining the sum of the probability of the sentence for all N-gram counts in the text corpus. The higher value of the score reflects that the sentence has correct grammar and structure. In addition to the language model correction, the author integrates a supervised spelling correction model similar to that proposed by Jinxi Guo et.al. [47].

VIII. INDONESIAN LANGUAGE SPEECH DATASET

In this section, we explain in detail the build process of the dataset used for training and evaluation of the system. The dataset is a result of combination between several available corpus, internet resources and data collected privately from Bahasa Indonesia Speaker.

A. DATA SOURCE

For this research, a total of 29,828 Bahasa Indonesia speech utterance was collected from various sources. This collection comprises television news broadcast, movie and drama transcript, Internet voice calls, and book recitations totaling in 27.5 hours of speech. The news broadcast data were taken from Metrotvnews [48], Tvonenews [49], and CNN Indonesia channel [50]. The movie and drama transcript is taken from the Bahasa Indonesia Corpus created by Elok Cahyaningtyas and Dhany Arifianto [15] While the book recitations data are recorded personally in the anechoic chamber. The speech data has an average length of 5 seconds for each sentence. For longer speeches, segmentation was completed to split them into two 5 second segments. Word labels were manually generated for speeches lacking annotation until the entire dataset had its own word labels. All speech included in the dataset are recorded in single channel ensured to be phonetically balanced. Figure 16 shows the distribution of the content in the speech dataset.

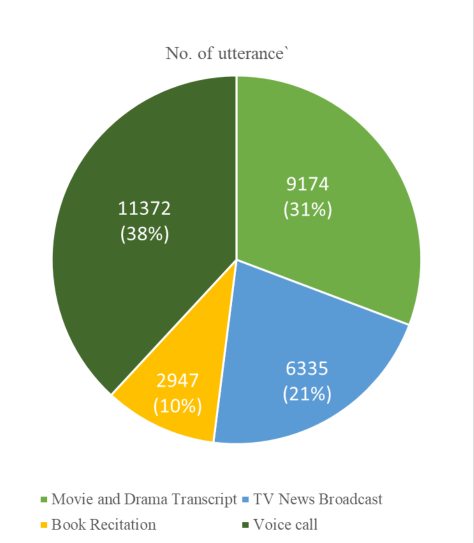

FIGURE 16. Distribution of speech data content in the dataset.

B. SPEAKER PROFILE

A total of 55 people which consist of 31 male and 24 female contributes to the speech dataset that is described above. The speaker age ranges from 16 to 44 years old with various background. For the television news data, the speaker is professional news caster and the movie and drama script are spoken professional announcer as described in [15]. The book recitation and voice call are performed by amateur with no experience in professional field.

While we tried to capture the various accent that Bahasa Indonesia speaker has, our current dataset has limited range of accent which mainly dominated by Java and Sundanese accent. However, because most of the sentence is spoken in formal way, the accent bears minimal effect to the spoken word. Figure17, 18, and 19 shows the overall speaker's profile for the dataset.

C. SPEAKER PROFILE

While some parts of the speech dataset are recorded in 16khz sample rate, the voice call which is a big part of the dataset was recorded in 8khz sampling rate.to assess this problem we considered three options to evaluate. The first one is to downsample all the speech recording to 8khz before used in the training process. This process will degrade the quality of the recording that was originally recorded in higher bitrate but maintains the uniformity and phonetically balanced nature of the dataset. This process also eases the computation load of the system, making the training and decoding process faster.

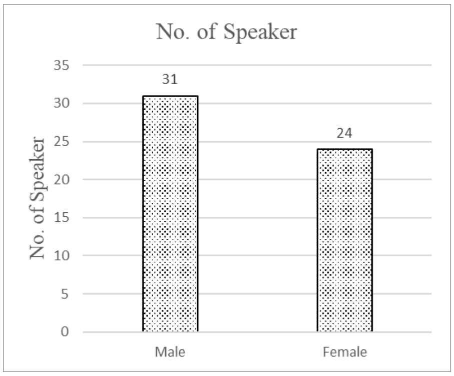

FIGURE 17. Gender distribution of the speakers.

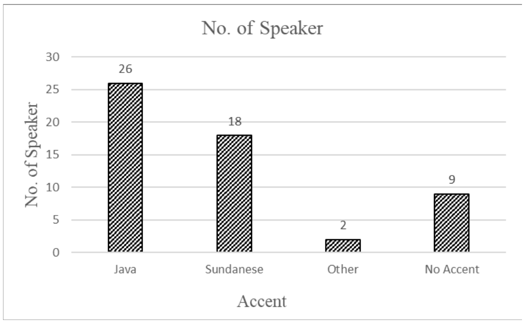

FIGURE 18. Accent distribution of the speakers.

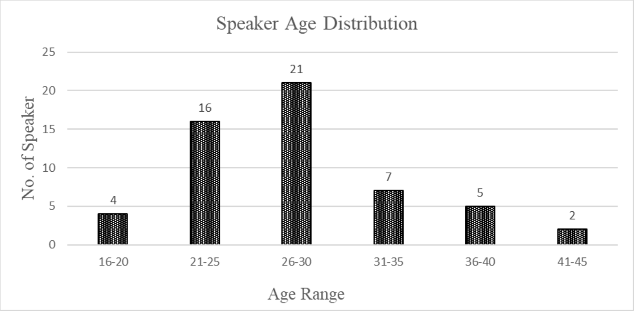

FIGURE 19. Age distribution of the speakers.

The second option is by discarding completely the voice call part of the dataset and keeping only recording with 16khz sampling rate. While this option does not degrade the speech quality large part of the vocabulary and information is discarded and the overall distribution in the dataset is heavily affected.

The last option to upsample the voice call dataset into 16khz. Similar to the first option, this process ensure that the

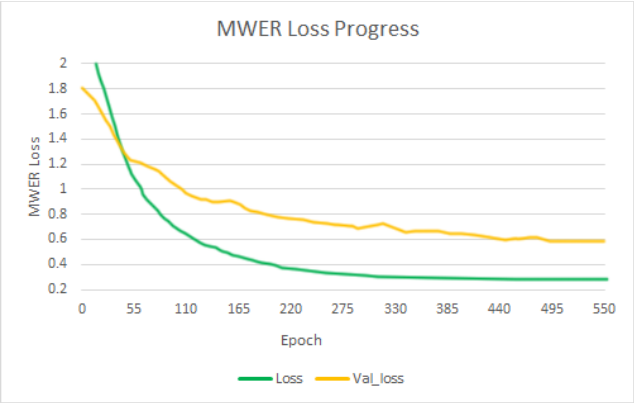

FIGURE 20. MWER loss progress throughout the training of transcript generation model.

TABLE 7. Dataset options.

Dataset Option	Speech Volume (Utterances)	Vocabulary Size	Phonetically Balanced
8Khz downsampled **(8DS)**	29828	31784	Yes
16khz No Voice Call **(16NVC)**	18456	12993	No
16khz Upsampled **(16UP)**	29828	31784	Yes

uniformity and the balance of the speech dataset is retained. However, upsampling the voice call part of the dataset does change the original quality of the recording nor add any information to it. Moreover, processing the additional unnecessary data will significantly slow the computation time. The detail for each dataset option is shown in Table 7. To find the optimal dataset option we evaluate it with three different model, we explore the evaluation process and result in section VII, Evaluation.

IX. EVALUATION

In this section, we conduct a series of evaluation scenario to observe the performance of each component under various condition. To evaluate the proposed method, we mainly used word error rate (WER) and word accuracy which is widely used as the assessment metric for word and text recognition. Word error rate indicates how well a model can predict the overall word in the transcript by comparing the edit distance of the predicted word to the actual word in the ground truth. A smaller WER indicates better performance of the model at predicting words. Word accuracy is a ratio of the correctly detected words in the transcript label divided by the total correct words in the ground truth label. The edit distance consists of a minimum number of substitutions, insertions, or deletions needed to render two words identical. This equation describes the WER calculation:

$$WER = \frac{S + D + I}{N} \tag{16}$$

Here, S is the number of substitutions, D is the number of deletions, and I is the number of insertions. A smaller WER indicates better performance of the model at predicting words. Word accuracy is a ratio of the correctly detected words in the transcript label divided by the total correct words in the ground truth label. Each evaluation scenario is detailed in the following subsection.

TABLE 8. Dataset options.

Dataset Option	Size (GB)	Training Time (Hour)
(8DS)	6.37	108
(16NVC)	6.84	131
(16UP)	12.73	294

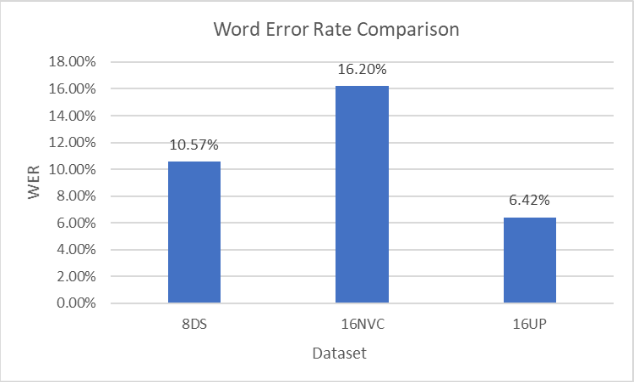

FIGURE 21. Transcript generation architecture based on LAS model it consists of 3 parts interconnected to each other.

A. DATASET EVALUATION

In this evaluation we train the transcript generation model with several dataset option as described in previous section and compare its training progress and performance in single speech situation. This evaluation helps us to gauge the trade-off between speech quality and speech volume and its impact again system performance and computation load. The training process for each dataset is identical to the one described in IV. D. Table 8 shows the comparison of dataset size and training time.

To compare the performance of each dataset we record new speech utterance from 6 speakers (3 male and 3 female).

These speakers are part of the original dataset while the recorded sentence is not. 15 speech utterance was recorded from each speaker totaling in 90 utterance used for evaluation. Transcript post-processing is not used for this test. We calculate WER from each model for comparison. Figure 21 shows the performance for each model.

As expected, our result shows that the model trained with 16UP, complete speech dataset sampled in 16khz give the best result of 6.42% this is the result of the wide-band speech quality that is preserved and the contextual information from the voice call dataset. However, training with this dataset takes more than twice the time of another dataset. On the other hand, 16NVC dataset perform the worst with 16.20% WER. Having no voice call dataset which compromise majority of the speech dataset degrade the performance significantly even if the speech is sampled in 16khz. this performance reflects the importance of the speech data volume for this model. Lastly 8DS dataset that is treated with downsample process perform 4.15% worse compared to the 16UP dataset, this performance gap is caused by the degradation of the speech signal from the downsampling. When factoring the time needed to train, model using 8DS dataset comes at the top, requiring only 36% the time it needed to training 16UP. Considering some system limitation on memory and computing capability, this could prove a feasible alternative dataset to train with.

Moving forward to other evaluation on the system, we consider using model trained 16UP dataset which perform best as the default when comparing with other variables.

B. MODEL PERFORMANCE COMPARISON

In this evaluation, we compare the transcript generation performance of the proposed method against several conventional methods mainly used in the Bahasa Indonesia speech recognition such as phoneme-based Sphinx4 ASR system adapted from research by Rifqy and Fanany [8], ASR system using English-based acoustic model (EBA model) by Veri and Ayu [51], and ASR for Bahasa Indonesia using CMUSphinx [52]. In addition, we also choose Deep-Speech [53] architecture that also use deep learning approach to compare against.

In the first task, we use the same testing dataset as in previous evaluation that consist of 90 clean speech utterance. Each model is assigned to generate transcript for this single-speaker speech and the WER and word accuracy for the transcript is calculated. Table 9 summarizes the result obtained from this task.

TABLE 9. Model performance comparison.

Method	WER (%)	Word Accuracy (%)
Sphinx-4 Based	24.15%	75.17%
EBA Model	29.66%	79.40%
CMU Sphinx	19.95%	74.64%
Proposed	11.16%	90.20%
Deep Speech	22.43%	72.80%
Proposed w/ Post Processing	5.25%	92.70%

As Table 9 demonstrates, compared to other methods, the proposed method has proven to have the lowest WER and highest word accuracy. Furthermore, with the addition of post-processing for the transcript, the accuracy can be further increased resulting in 5.25% WER. The results of the different methods indicate that the other method struggles with detecting words with a fast pronunciation and specific accent, while the proposed model from the speech signal, can detect words more accurately regardless of the speech variance.

Surprisingly, in this task deep speech architecture that we adapt perform worse compared to the CMU sphinx method with a difference of 2.48% WER. We suspect this is due to the composition of deep speech architecture that consist of deep network that requires a lot of training data to capture the speech context which is suitable in its original application. However, as we only have limited training dataset the performance of the model is severely constrained.

In the second task, we choose one female speaker and one male speaker from the previous training dataset and combine the speech resulting in 10 M-F mixture utterance. We then, use speech separation and transcript each individual speech and calculate its WER. As the previous models used for comparison does not have any speech separation module, we use ours for every case. The result is shown in Table 10.

TABLE 10. M-F Mixture performance comparison.

Method	WER-F (%)	WER-M (%)
Sphinx-4 Based	33.95%	29.53%
EBA Model	32.18%	34.85%
CMU Sphinx	21.93%	22.68%
Proposed	15.34%	17.38%
Deep Speech	23.83%	23.44%
Proposed w/ Post Processing	7.86%	9.82%

From the result, we could observer that the performance across the model is consistent with the previous task, however there are overall decrease in the performance due to the separated speech that is slightly degraded and is not possible to be reconstructed perfectly. The performance of each gender varies between models within less than 3% of difference.

C. SPEECH SEPARATION EVALUATION

In this section we will shift our focus on evaluating the performance of the model when it is used in the concurrent speech condition. For this evaluation, we use our proposed model with post-processing applied to the transcript.

First, we observe the performance of the system in 3-person concurrent speech situation. We use the same speaker and dataset in the previous evaluation and combine their speech to form every permutation. The system then generates the speech transcript for every separation target. Finally, WER is calculated for each transcript. Table 11 summarize the of this experiment.

TABLE 11. System performance in 3-person speech situation.

Separation Target	Mixture			
	FFF	MMM	FMM	FFM
F1	17.22%	-	13.77%	13.38%
F2	15.37%	-	12.44%	18.71%
F3	18.58%	-	12.84%	15.31%
M1	-	20.69%	15.79%	16.89%
M2	-	14.23%	15.78%	15.78%
M3	-	15.29%	14.57%	12.84%

As indicated in several related research, the mixture of gender plays a significant role in the performance of the system. Generally, a mixture from the same gender is harder to separate which result in worse performance. This is due to the similarity of the pitch between the speakers. Separating a different gender from a 3-person mixture (e.g., separating male speaker from FFM) results in better performance as the male speaker have more distinct pitch contour compared to the two female speakers. Despite this nature, we could observe the difference of separating different gender (FF or MM) with opposite gender (FM) is quite small in this task owing to the ability of the system to track the pitch of the speaker in the speech separation process.

In the next task, a mixed speech from multiple simultaneous speakers was used. The experiment was performed for situations with two, three, and four speakers mixed. For each situation, a speaker was chosen from the combination at random as a separation target. The test was conducted 20 times for each situation and calculated the minimum, average, and maximum WER for each situation. Table 12 illustrates the results of this test case.

TABLE 12. Performance comparison with combined speech.

No. of Speaker	Min. WER (%)	Avg. WER (%)	Max WER (%)
2	6.55%	8.00%	10.99%
3	10.57%	14.50%	20.74%
4	28.27%	31.20%	40.15%

As Table 2 indicates, the addition of speakers in the combined speech decreased the overall performance of the system due to the additional speech separation process needed in the combined speech situation. The two-speaker situation demonstrated little performance reduction over the single-speaker situation. The worst performance occurred in the four-speaker situation, where the average WER was 31.20%, with the worst-performing singular case having an average WER of 40.15%.

D. SPEAKER IDENTIFICATION EVALUATION

As a part of the system, speaker identification module does not directly affect the quality of the generated transcript. However, it improves the system usability and saving extra manual step for the user to assign the written transcript to each speaker. We evaluate the speaker identification module by giving the following task: The module attempts to identify a single speaker from the available 2 speakers to 6 speakers. These speakers are chosen from testing set we use in the previous evaluation. For each guess the chosen speaker is randomized and 10 guesses are conducted for each number of people (10 guesses for 2 speakers, 10 guesses for 3 speakers and so on) The identification result is shown in Figure 22.

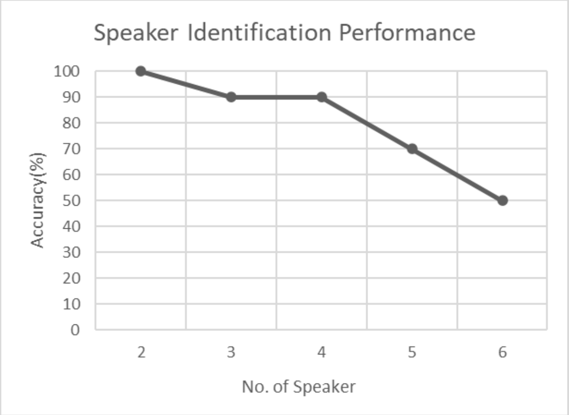

FIGURE 22. Trend of speaker identification performance with the addition of speaker.

It could be seen from the figure that in a situation where there are only two speakers, the module could perfectly identify each of the speakers. However, the trend shows a significant drop when the module attempts to identify a speaker among 5 speakers. In the most extreme test case where there are 6 speakers. The module was only able to identify the speakers 50% of the time. From this result, we conclude that the speaker identification module could be practically used for the situation where there are no more than 4 speakers.

E. ADAPTATION WITH OTHER LOW-RESOURCE LANGUANGE

Including Bahasa Indonesia, there are many low-resource languages around the world that is not sufficiently explored. In this experiment we attempt to train our model with several others low-resource language and measure how well it performs in generating transcript. For this experiment we choose Basque, Malay and Afrikaans language as comparison because it shares a same or similar character grapheme with Bahasa Indonesia as opposed the likes of Bengal language which have character on their own. The open speech dataset for these languages is also readily available which made it easier to process and control the training variable.

For Basque, we use Common Voice Basque speech dataset [54] that contains 65 hours validated speech from 638 speakers. 55 hours of the speech is used for training where 10 hours of speech is dedicated for testing. For Malay language, MASS: A Malay language LVCSR corpus [55] is chosen as dataset. 70 hours of read speech is available where, similar to Basque, we use 60 hours of it for training dataset and 10 hours for testing. Lastly, for Afrikaans language, the combination of NCHLT Afrikaans corpus [56] and Lwazi II Afrikaans Trajectory Tracking Corpus [57] is used. NCHLT corpus consists of 56 hours speech spoken by 210 participants while Lwazi II Afrikaans Trajectory Tracking Corpus contains 4 hours of speech by single speaker. In total 60 hours of speech are obtained, where 50 hours of speeches are used for training and 10 hours of speeches are used for testing.

Compared to the previous experiment architecture of the system is unchanged. However, no speech separation and post-processing is done to the transcript as currently we do not have access to the language model of Afrikaans, Malay and Bisque. All speech used for training and testing is first enhanced with Iterative Kalman filter and we use the same input feature across the language. In order to minimize the discrepancy between training dataset, we use 16NVC dataset that contains less speech data instead of 16UP.

For testing, the system attempts to generate 10 hours of testing dataset from respective language. The WER for each transcript is calculated and the result is shown in Figure 22.

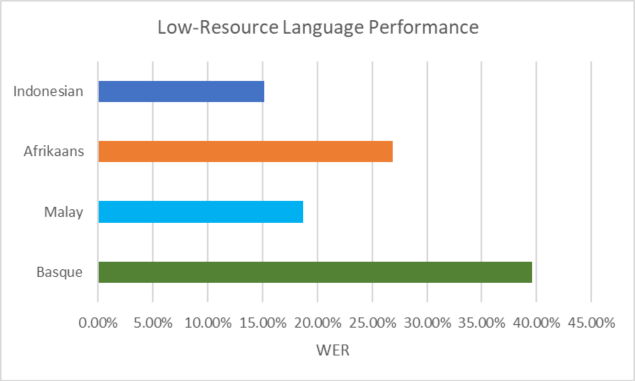

FIGURE 23. Comparison of the system, trained with various low-resource languages on 10 hour of testing dataset.

Our observation on the system performance shows a decent result on Afrikaans language with 26.91% WER. The performance on Malay language with 18.7% WER is to be expected considering it is the root language of Bahasa Indonesia, sharing significant amount of word and similar structure. The performance is only 3.5% apart from Bahasa Indonesia which is used as baseline. Basque language come last in performance with 39.6% WER which is a substantial gap of 24.4% WER when compared to Bahasa Indonesia.

We believe the difference of the sentence structure and grammar contributes to this poor performance, while Afrikaans language also bears different structure with Bahasa Indonesia it still shares the common similarity of having a lot of compound words which explain the slightly better performance.

From this experiment we conclude that the proposed method is still applicable for another low-resource language with similar nature or language structure as Bahasa Indonesia. This could be further extended to other low-resource language by adjusting the parameter of the network and improving the volume of the training dataset, as the current experiment only utilize a limited amount of speech. A change in the grapheme sequence matching might also be required for languages that uses non-alphabetic character for their writing system.

X. CONCLUSION

This paper proposes a novel speech transcription generation method in Bahasa Indonesia that uses the combination of speech separation, an LAS network, and transcript processing. As a relatively low-resource language that is seldom explored, Bahasa Indonesia has a disadvantage in terms of communication-assistive technology development, and the proposed method could be useful in situations that involve simultaneous speech, such as online discussions and remote conferences. The proposed method uses pitch-aware speech separation to separate several speeches that are mixed in a single channel. A LAS based network is then used to generate a transcript. After the transcript has been generated, post-processing is completed with the help of the WordNet semantic network to enhance the accuracy of the system.

We have thoroughly explored the capability of the system by conducting a various type of evaluation. According to our dataset evaluation, voice-call dataset proven to be really important part to introduce speech context information to our system. When compared to other methods in a single speech and M-F mixture speech situation our system also shows to have the best performance with WER as low as 5.25% for single speech and 7.86% for mixture speech.

In situation with three speakers our system manages to keep the performance below 21% WER and is still capable to generate transcript for 4-speakers separation albeit with degraded performance. To proof the generality of our system we attempt to train the system with several different low-resource language which results in great performance in Malay language with 18.7% WER and acceptable performance in Afrikaans with 26.91%WER..

REFERENCES

[1] *Penduduk Indonesia Hasil Sensus Penduduk 2010 (Result of Indonesia Population Census 2010)*, Badan Pusat Statistik, Central Jakarta, Indonesia, 2013, pp. 421–427.

[2] T. Johannes. (2005). *Bahasa Indonesia, Between FAQs and Facts*. Accessed: Feb. 24, 2021. [Online]. Available: http://www.indotransnet.com/article1.html

[3] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, ''Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project,'' in *Proc. Workshop Technol. Corpora Asia–Pacific Speech Transl. (TCAST)*, 2008, pp. 1–6.

[4] A. Setyadi, ''The type of Indonesian language phoneme in 'minimal pair,''' in *Proc. 1st Int. Conf. Culture, Literature, Lang. Maintenance Shift (CLLAMAS)*, Semarang, Indonesia, Aug. 2019, p. 61.

[5] M. Muljono, A. Q. Syadida, D. R. I. M. Setiadi, and A. Setyono, ''Sphinx4 for Indonesian continuous speech recognition system,'' in *Proc. Int. Seminar Appl. Technol. Inf. Commun. (iSemantic)*, Oct. 2017, pp. 264–267.

[6] C. H. Satriawan and D. P. Lestari, "Feature-based noise robust speech recognition on an Indonesian language automatic speech recognition system," in *Proc. Int. Conf. Electr. Eng. Comput. Sci. (ICEECS)*, Nov. 2014, pp. 42–46.

[7] B. Distiawan and R. Manurung, "A GrAF-compliant Indonesian speech recognition Web service on the language grid for transcription crowdsourcing," in *Proc. 6th Linguistic Annotation Workshop*, Jeju, Republic of Korea, 2012, pp. 67–74.

[8] R. Adiwidjaja and M. I. Fanany, "End-to-end Indonesian speech recognition with convolutional and gated recurrent units," *J. Phys., Conf. Ser.*, vol. 1566, Jun. 2020, Art. no. 012118, doi: 10.1088/1742-6596/1566/1/012118.

[9] B. T. Atmaja, D. Arifianto, and M. Akagi, "Speech recognition on Indonesian language by using time delay neural network," in *Proc. ASJ Spring Meet*, 2018, pp. 1291–1294.

[10] B. T. Atmaja and M. Akagi, "Speech emotion recognition based on speech segment using LSTM with attention model," in *Proc. IEEE Int. Conf. Signals Syst. (ICSigSys)*, Bandung, Indonesia, Jul. 2019, pp. 40–44.

[11] J. J. Lasiman and D. Puji Lestari, "Speech emotion recognition for Indonesian language using long short-term memory," in *Proc. Int. Conf. Comput., Control, Informat. its Appl. (IC3INA)*, Tangerang, Indonesia, Nov. 2018, pp. 40–43.

[12] C. Anindya, D. Purwanto, and D. I. Ricoida, "Development of Indonesian speech recognition with deep neural network for robotic command," in *Proc. ISITIA*, Surabaya, Indonesia, Aug. 2019, pp. 434–438.

[13] E. R. Swedia, A. B. Mutiara, M. Subali, and Ernastuti, "Deep learning long-short term memory (LSTM) for indonesian speech digit recognition using LPC and MFCC feature," in *Proc. 3rd Int. Conf. Informat. Comput. (ICIC)*, Palembang, Indonesia, Oct. 2018, pp. 1–5, doi: 10.1109/IAC.2018.8780566.

[14] (2021). *NII-SRC Speech Resource Consortium, TITML-IDN—Speech Resources Consortium*. Accessed: Feb. 24, 2020. [Online]. Available: http://research.nii.ac.jp/src/en/TITML-IDN.html

[15] E. Cahyaningtyas and D. Arifianto, "Development of under-resourced Bahasa Indonesia speech corpus," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 1097–1101, doi: 10.1109/apsipa.2017.8282191.

[16] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 1992, pp. 517–520, doi: 10.1109/icassp.1992.225858.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210, doi: 10.1109/icassp.2015.7178964.

[18] X. Zhu and H. Huang, "End-to-end Amdo–Tibetan speech recognition based on knowledge transfer," *IEEE Access*, vol. 8, pp. 170991–171000, 2020, doi: 10.1109/ACCESS.2020.3023783.

[19] W. Zhang, H. Yang, X. Bu, and L. Wang, "Deep learning for Mandarin–Tibetan cross-lingual speech synthesis," *IEEE Access*, vol. 7, pp. 167884–167894, 2019, doi: 10.1109/ACCESS.2019.2954342.

[20] D. Liu, J. Xu, P. Zhang, and Y. Yan, "Investigation of knowledge transfer approaches to improve the acoustic modeling of Vietnamese ASR system," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 5, pp. 1187–1195, Sep. 2019, doi: 10.1109/JAS.2019.1911693.

[21] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Language-adversarial transfer learning for low-resource speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 621–630, Mar. 2019.

[22] C.-T. Lin, Y.-R. Wang, S.-H. Chen, and Y.-F. Liao, "A preliminary study on cross-language knowledge transfer for low-resource Taiwanese Mandarin ASR," in *Proc. Conf. Oriental Chapter Int. Committee Coordination Standardization Speech Databases Assessment Techn. (O-COCOSDA)*, Bali, Indonesia, Oct. 2016, pp. 33–38.

[23] F. Bond, L. T. Lim, E. K. Tang, and H. Riza, "The combined Wordnet Bahasa," *NUSA, Linguistic Stud. Lang. Around Indonesia*, vol. 57, pp. 83–100, Sep. 2014.

[24] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979, doi: 10.1109/proc.1979.11540.

[25] S. Ayat, M. T. Manzuri, R. Dianat, and J. Kabudian, "An improved spectral subtraction speech enhancement system by using an adaptive spectral estimator," in *Proc. Can. Conf. Electr. Comput. Eng.*, 2005, pp. 261–264.

[26] N. Cheng, W.-J. Liu, and B. Xu, "An improved a priori MMSE spectral subtraction method for speech enhancement," in *Proc. 3rd Int. Workshop Signal Design Appl. Commun.*, Sep. 2007, pp. 373–377.

[27] Q. Li, R. Li, K. Ji, and W. Dai, "Kalman filter and its application," in *Proc. 8th Int. Conf. Intell. Netw. Intell. Syst. (ICINIS)*, Nov. 2015, pp. 74–77, doi: 10.1109/ICINIS.2015.35.

[28] M. Andra and T. Usagawa, "Single channel speech enhancement for deep convolutional neural network feature using EM-Kalman filter," in *Proc. Acoust. Soc. Japan Student Meeting Kyuushu*, 2017, pp. 67–70.

[29] G. Sharon, "Speech enhancement: Application of the Kalman filter in the estimate-maximize (EM) framework," in *Speech Enhancement* (Signals and Communication Technology). Berlin, Germany: Springer, 2005.

[30] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, nos. 7–8, pp. 588–601, Jul. 2007, doi: 10.1016/j.specom.2006.12.006.

[31] I. Jafari, R. Togneri, and S. Nordholm, "Review of multi-channel source separation in realistic environments," in *Proc. 13th Australas. Int. Conf. Speech Sci. Technol.*, 2010, pp. 201–204.

[32] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 31–35, doi: 10.1109/ICASSP.2016.7471631.

[33] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 241–245, doi: 10.1109/ICASSP.2017.7952154.

[34] Y. Liu and D. Wang, "A casa approach to deep learning based speaker-independent co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5399–5403.

[35] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 61–65, doi: 10.1109/ICASSP.2017.7952118.

[36] K. Wang, F. Soong, and L. Xie, "A pitch-aware approach to single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 296–300.

[37] H.-T. Luong and J. Yamagishi, "A unified speaker adaptation method for speech synthesis using transcribed and untranscribed speech with backpropagation," 2019, *arXiv:1906.07414*. [Online]. Available: http://arxiv.org/abs/1906.07414

[38] M. Seurin, F. Strub, P. Preux, and O. Pietquin, "A machine of few words—Interactive speaker recognition with reinforcement learning," Aug. 2020, *arXiv:2008.03127*. [Online]. Available: http://arxiv.org/abs/2008.03127

[39] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," NIST, Gaithersburg, MD, USA, Tech. Rep. LDC93S1, 1993.

[40] S. Sakti, A. A. Arman, S. Nakamura, and P. Hutagaol, "Indonesian speech recognition for hearing and speaking impaired people," in *Proc. INTERSPEECH*, 2004, pp. 1037–1040.

[41] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 4960–4964, doi: 10.1109/ICASSP.2016.7472621.

[42] J. Koutnik, K. Greff, F. Gomez, and J. Schmidhuber, "A clockwork RNN," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1863–1871.

[43] M. B. Andra and T. Usagawa, "Automatic lecture video content summarizationwith attention-based recurrent neural network," in *Proc. Int. Conf. Artif. Intell. Inf. Technol. (ICAIIT)*, Yogyakarta, Indonesia, Mar. 2019, pp. 54–59.

[44] F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara, "Revising perceptual linear prediction (PLP)," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 1–4.

[45] M. Magimai-Doss, "Phoneme vs grapheme based automatic speech recognition," Idiap Res. Inst., Martigny, Switzerland, Tech. Rep. Idiap-RR-48-2004, 2004.

[46] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 4774–4778, doi: 10.1109/ICASSP.2018.8462105.

[47] J. Guo, T. N. Sainath, and R. J. Weiss, "A spelling correction model for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 5651–5655.

[48] YouTube. (2020). *Metrotvnews*. Accessed: May 24, 2020. [Online]. Available: https://www.youtube.com/user/metrotvnews.

[49] YouTube. (2020). *tvOneNews*. Accessed: May 24, 2020. [Online]. Available: https://www.youtube.com/channel/UCER4rvDnRBPr_ncYW4UCZjg

[50] YouTube. (2020). *CNN Indonesia*. Accessed: May 24, 2020. [Online]. Available: https://www.youtube.com/user/CNNindonesia

[51] V. Ferdiansyah and A. Purwarianti, "Indonesian automatic speech recognition system using English-based acoustic model," in *Proc. Int. Conf. Electr. Eng. Informat.*, Bandung, Indonesia, Jul. 2011, pp. 1–4.

[52] H. Prakoso, R. Ferdiana, and R. Hartanto, "Indonesian automatic speech recognition system using CMUSphinx toolkit and limited dataset," in *Proc. Int. Symp. Electron. Smart Devices (ISESD)*, Bandung, Indonesia, Nov. 2016, pp. 283–286.

[53] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.

[54] A. Rosana, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2019, *arXiv:1912.06670*. [Online]. Available: https://arxiv.org/abs/1912.06670

[55] T.-P. Tan, X. Xiao, E. K. Tang, E. S. Chng, and H. Li, "MASS: A Malay language LVCSR corpus resource," in *Proc. Oriental COCOSDA Int. Conf. Speech Database Assessments*, Urumqi, China, Aug. 2009, pp. 25–30.

[56] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proc. SLTU*, Saint Petersburg, Russia, May 2014, pp. 1–7.

[57] J. Badenhorst and M. Davel. (2021). *Lwazi II Afrikaans Trajectory Tracking Corpus*. Accessed: Feb. 11, 2021. [Online]. Available: https://hdl.handle.net/20.500.12185/442

MUHAMMAD BAGUS ANDRA (Member, IEEE) received the B.E. degree from the Institut Teknologi Sepuluh Nopember, in 2015, and the M.S. degree in computer science and electrical engineering from Kumamoto University, where he is currently pursuing the Ph.D. degree in computer science. He is a member of the Human Interface Cyber Communication Laboratory. His research interests include speech processing, natural language processing, and machine learning. He is also a member of ASJ.

TSUYOSHI USAGAWA (Member, IEEE) received the B.E. degree from the Kyushu Institute of Technology, in 1981, and the M.E. degree from Tohoku University, in 1983. Since 1983, he has been with Kumamoto University, where he has been a Professor with the Graduate School of Science and Technology, since 2004. He is a member of ASA, ASJ, INCE/J, IEICE, JSET, JAIS, and ACM. From 2005 to 2007, he was the Vice President of ASJ. From 2014 to 2020, he was the Vice Dean with the Graduate School of Science and Technology, Kumamoto University. He is currently acting as the Trustee Vice President with Kumamoto University. His research interests include acoustic signal processing, perceptual information processing, educational data mining, and e-learning among others.