Combining Fuzzy Clustering and Hidden Markov Models for Sundanese Speech Recognition

Intan NurmaYulita, Akik Hidayat, Atje Setiawan Abdullah, Erick Paulus
Departement of Computer Science, UniversitasPadjadjaran, Jl. Raya Bandung – Sumedang Km. 21, Sumedang 45363, Indonesia

intan.nurma@unpad.ac.id

Abstract. Sundanese tribe is one of the largest population tribe in Indonesia. However, over time, users of the Sundanese language are declining because of the living languages outside of Sundanese. One way to preserve Sundanese is Sundanese Speech Recognition. In this research, several processes of recognition were done include pre-processing, feature extraction, Fuzzy Clustering, and Hidden Markov Models. Pre-processing aims to separate the recording from the noise and normalize the speech signal, while the feature extraction to obtain the characteristics of the speech signal to distinguish each phoneme from the speech. In particular, the contribution of this research is to combine Fuzzy Clustering and Hidden Markov Models for Sundanese Speech Recognition. Fuzzy Clustering plays a role in finding unique symbols in the speech signal. These symbols are represented as centroid in fuzzy clustering. The next process, each segment of the speech signal calculated the probability of the membership for all centroids. The output of this calculation becomes input to Hidden Markov Models. The test uses a speech corpus derived from 30 people. The results obtained that the combination of Fuzzy Clustering and Hidden Markov Models have a better performance than Hidden Markov Models. Also, the research also analyses the optimal number of clusters of Fuzzy Clustering and states of Hidden Markov Models for the datasets used.

1. Introduction
Java island, one of the largest islands of Indonesia [1]. Sundanese language is used in everyday conversations as well as a Bandung government program to use the Sundanese language on Wednesday. Sundanese cultivation is evidenced by the method of introduction of Sundanese language with the game [2]. Sundanese language also contributes significantly to the development of Indonesian language as a national language. There is much vocabulary that is absorbed into the Indonesian language which every year Indonesian language develops its vocabulary from the Sundanese language. It is because the Sundanese language has more vocabulary and has three levels of language type used. Thus, Sundanese language becomes one of the critical languages in Indonesia that needs to be preserved.

One way to preserve what can be done is through learning Sundanese language. However Sundanese language teachers are still very minimal, so it takes learning of Sundanese language through self-learning. This mechanism can be done through an application that presents speech recognition. Of course, future speech recognition will be beneficial in the area of speech technology widely.

Many studies have been developed for speech recognition, such as Convolutional Neural Networks [3] [4], Decision Tree [5], Deep Belief Networks [6], Recurrent Neural Network [7]. However, the most widely used method is Hidden Markov Models (HMM). This method consists of two main stages
of quantization and model development. Quantization is done through clustering, and the model is built using a probabilistic approach. The prediction of a model is based on the highest probability value between the data to be predicted with all previously recognized models.

The challenge in speech recognition is the same word or phrase even though spoken by the same speaker will always produce a different signal form. HMM cannot appropriately handle the uncertainty of this nature. The most widely used method for this problem is Fuzzy Logic. Therefore, a method that can combine the advantages of both methods for Speech Recognition is required. Several studies have combined both methods [8] [9]. One method that has applied it to Speech Recognition is Fuzzy Hidden Markov Models (FHMM) developed by IN Yulita et al. [10].

However, the research has not been developed for Sundanese. Therefore, this study aims to apply the same method to Sundanese where Fuzzy Clustering is applied as a machine-generated feature extraction, and FHMM receives input from the learning feature to build the optimal model.

2. Related works
The three main parts in speech recognition are feature extraction, clustering, and classifier.

2.1 Linear Predictive Coding
Linear Predictive Coding (LPC) is commonly used in audio processing activities especially in the speech recognition. By using a linear model, this technology will present a correlated data property. The higher the value will affect the high-frequency resolution [11]. The use of LPC is illustrated by the following formula [12].

$$\hat{x}_M(n) = - \sum_{i=1}^{M} b_{M,i} x(n - i)$$  

(1)

In Equation 1,$\hat{x}_M(n)$is the random process of the coefficient $M$ where it is a part of a predictive filter for working on the accuracy of the prediction. $b_{M,i}$is the best predicted taken from the data sample. $x(n-i)$ is a sample of the previous value. The next calculation is the error prediction of the LPC by using the difference formula between the target and the result of the calculation.

$$e_M(n) = x(n) - \hat{x}_M(n) = x(n) + \sum_{i=1}^{M} b_{M,i} x(n - i)$$  

(2)

$\hat{x}_M(n)$is a representation of the incoming signals at time n. Due to the absence of new information on the target component of prediction calculation then the formula becomes simple.

$$H_M(z) = 1 + \sum_{i=1}^{M} b_{M,i} z^{-i}$$  

(3)

The variance value of the LPC with input $x(n)$ of the target prediction error is as follows.

$$\sigma^2 = E[|e_M(n)|^2] = \sum_{i=1}^{M} [e_M(n) - \hat{e}_M]^2$$  

(4)

2.2 Fuzzy Clustering
Clustering aims to group data so that data with the same characteristics have the same group and data with different characteristics have different groups [13]. Fuzzy C-means Clustering (FCM) is one of the clustering methods that are part of the K-Means Clustering method. FCM uses a fuzzy grouping model so that data can be a member of all classes or clusters formed with different degrees or membership levels between 0 and 1 [14]. The degree of membership determines the degree of data presented in a class or cluster. The basic concept of FCM first is to determine the cluster center that will mark the average location for each cluster. In the initial conditions, the center of the cluster is still not accurate. Each data has a membership degree for each cluster. By repairing the cluster center and membership value of each data repeatedly, it can be seen that the centroid will approve the appropriate
location. This process is based on the minimization of objective functions that describes the distance from the data points assigned to the centroid of the cluster which is weighted by the degree of membership of that data.

2.3 Fuzzy Hidden Markov Models
Fuzzy Hidden Markov Models (FHMM) is a method developed from HMM by combining it with Fuzzy Clustering [10]. Specifically, input from FHMM is not a quantization result. Fuzzy Clustering as a pre-processing is used as a machine-generated feature extraction. It means that each segment that initially has features of the feature extraction is converted to a matrix with degrees of membership. However, the input of Fuzzy Hidden Markov Models is certainly cannot be from the membership degree matrix but also data with more than one feature. The models in FHMM are based on the following equations:

- Initialize on forwarding calculation
  \[ x_1 (i) = \pi_1 b_i (d_1) \]  

- Induction on forwarding calculation
  \[ x_{t+1} (j) = \sum_{j=1}^{N} x_t (i) x_{jl} b_j (d_{t+1}) \]  

- Induction on backward calculation
  \[ y_t (j) = \sum_{j=1}^{N} a_{jl} b_j (d_{t+1}) B(j) \]  

- Equation of final forward and backward
  \[ P(O|\lambda) = \sum_{i=1}^{N} \sum_{j=1}^{N} x_t (i) a_{jl} b_j (d_{t+1}) B_{t+1}(j) \]  

Where

- Fuzzy observation data
  \[ b_j (d_t) = \sum_{z=1}^{c} B_{zj} K_{tz} \]  

- Equation of membership degree
  \[ K_{tz} = \frac{\left( \frac{\sum_{i=1}^{m} (l_{if} - V_{zf})^2}{\sum_{i=1}^{m} l_{if} - V_{zf}} \right)^{-1}}{w^{-1}} \]  

3. Methodology
Implementation of the proposed methods was tested using the dataset of the architecture.

3.1 Dataset
The data set came from 10 men consisting of five men and five women who are Sundanese native speakers. The recording was done in the soundproof room to reduce the noise that may occur during the recording process. The recording took place in June 2017, and each speaker was in charge of pronouncing digits in Sundanese 60 times. Table 1 shows the list of Sundanese digits used in this study. The intercurrent pronunciation and repetition were stored in a wav file to streamline time during sound recording.

| No | Digit | Meaning |
|----|-------|---------|
| 1  | Enol  | Zero    |
| 2  | Hiji  | One     |
| 3  | Dua   | Two     |
| 4  | Tilu  | Three   |
| 5  | Opat  | Four    |
| 6  | Lima  | Five    |
| 7  | Genep | Six     |
| 8  | Tujuh | Seven   |
| 9  | Dalapan | Eight |
| 10 | Salapan | Nine  |

3.2 Architecture
Speech recognition conducted in this study consists of several stages, as shown in Figure 1.

![Figure1. Architecture of speech recognition](image)

3.2.1 Pre-processing
Pre-processing consisted of four processes: segmentation, centering, normalization, and windowing. Segmentation was done because all the list of words and repetitions were stored in the file. Therefore, segmentation is required to trim the file into word units so that one file holds only one word. This process was done manually and the longest time in speech recognition in this research. The result of segmentation was then processed with centering that aimed to shift the location of the discrete amplitude to be in the middle. The next process was normalization that aims to equalize the maximum amplitude of all sound signals. The final process of pre-processing was windowing which divides each sound file by a particular frame size. Windowing was done overlapping to avoid missing information.
at the break between frames. The result was that each sound file has some frames/segments. The longer a file, the more frames/segments it had. It allowed the same word spoken by the same speaker to have a different number of frames.

3.2.2 Feature extraction
Feature extraction in this research was LPC. The output of LPC generated 24 features. The feature extraction mechanism at this stage is called hand-crafted feature extraction. If a file consisted of 25 frames/segments, then this stage produced a 25 * 24 matrix.

3.2.3 Clustering
Clustering in this study had two tasks that form the codebook (center cluster) and machine-generated feature extraction. In the machine-generated feature extraction stage, the 24 features in the previous stage calculated the degree of membership to all the codebooks. The output of this stage was, each sound file with 25 frames will have a matrix of 25 * number of clusters. If clustering produces 16 clusters, then the matrix constructed is 25 * 16.

3.2.4 Pattern Recognition
Learning phase in this research built a model for each word that will be known as many as 10 model. At the testing stage, the predicted sound signal counted its chances towards the entire model. The model that has the highest chance (maximum likelihood) becomes predicted label.

3.2.5 Evaluation
Testing was done by dividing the data into two parts that were 70% of the dataset as training data while the rest of test data. The evaluation parameters in this study based on accuracy, precision, recall, and F-measure.

4. Experiments
Some tests were performed on the corpus speech to find the optimal conditions for the parameters of the number of clusters, states, and membership degrees. Tests for analysis of the number of clusters and states were based on K-Means Clustering and HMM. Analysis of the influence of membership degree based on Fuzzy Clustering and Fuzzy Hidden Markov Models.

4.1 Effect Analysis of Number of Clusters
Experiments were conducted by testing the variation of the cluster number parameters of 16, 32, 64, 128 from K-Means Clustering. Codebook generated from each experiment is tasked to quantify the HMM input where the number of states used is five. Table 2 shows the experimental results of clusters analysis.

| Number of clusters | Accuracy | Precision | Recall | F-Measure |
|--------------------|----------|-----------|--------|-----------|
| 16                 | 8.00     | 10.00     | 0.96   | 1.76      |
| 32                 | 16.00    | 10.00     | 1.70   | 2.91      |
| 64                 | 16.00    | 10.00     | 1.70   | 2.91      |
| 128                | 16.00    | 10.00     | 1.70   | 2.91      |

Table 2 shows the lowest performance when used 16 clusters. It indicates that the phonemes of the compilers had not been represented in their entirety by using only. If the number of clusters was 32,
the system performance increased. However, adding more clusters to 64 and 128 did not have an impact because the system did not generate.

4.2. Effect Analysis of Number of State

Analysis the effect of the number of states by changed the number of states from five to ten, but the cluster parameters were constant. The number of clusters used in this experiment was 32 by the optimum results obtained in the experiments in subchapter 4.1. Table 3 describes the results of experiments on states in HMM.

| Number of states | Accuracy | Precision | Recall | F-Measure |
|------------------|----------|-----------|--------|-----------|
| 5                | 8.00     | 10.00     | 0.96   | 1.76      |
| 6                | 16.00    | 10.00     | 1.70   | 2.91      |
| 7                | 16.00    | 10.00     | 1.70   | 2.91      |
| 8                | 16.00    | 10.00     | 1.70   | 2.91      |
| 9                | 8.00     | 10.00     | 0.96   | 1.76      |
| 10               | 16.00    | 10.00     | 1.70   | 2.91      |

Based on Table 2, it can be seen that the increase of states did not always have an upward performance trend. The lowest performance occurred when the states used were five and nine. The number of states in the HMM had a role to capture the variations of sound units. If the states were too few but variations were much, then the results of the recognition to be small. In Table 3, states had been changed from five to nine, but the performance remained small. It means that state is not the primary cause of low performance.

4.3 Analysis of the effect of degree of membership

The highest performance on the previous is 2.91% for F-Measure. The next scenario was to implement it in FHMM by using the optimized parameters obtained in both tables. The analysis focuses on the effect of membership degrees. The results are shown in Table 4.

| Size of membership degree (w) | Accuracy | Precision | Recall | F-Measure |
|-------------------------------|----------|-----------|--------|-----------|
| 1.05                          | 85.40    | 87.97     | 87.05  | 86.63     |
| 1.1                           | 86.67    | 90.10     | 88.82  | 88.67     |
| 1.2                           | 89.53    | 92.03     | 90.77  | 91.11     |
| 1.3                           | 86.27    | 89.33     | 87.87  | 88.35     |
| 1.4                           | 82.67    | 85.99     | 85.09  | 84.87     |
| 1.5                           | 78.40    | 82.15     | 81.57  | 80.80     |

Based on Table 3, the highest performance was achieved when w = 1.2. The small value had a lower performance. It was due to the small cluster coverage that is unable to provide similarities between the sound units of other sound units. However, too big of w the performance became smaller, as it happened when w = 1.5. It was due to the cluster coverage that is too large the more unable to distinguish between sound units.
5. Conclusion

The experiments show the lowest conditions of HMM were achieved when the clusters were 16, and the states were five. Fewer clusters caused less system to store variations of sound units in the dataset. When compared to FHMM in Table 4, there was a significant increase in performance. It indicates that uncertainty in speech corpus needed to be addressed. In this research, implementation was done through Fuzzy HMM. The optimized condition was obtained when $w = 1.05$. If $w$ was too small, it caused the system could not see the similarity that occurred between the sound units, but if $w$ is too large, it caused the system was not able to distinguish between sound units well.

References

[1] HR Dhona 2017 Wilayah sunda dalam surat kabar sunda era colonial Jurnal Komunikasi 11(1) 1-16

[2] A Wahyuningsih dkk 2015 Meningkatkan kosakata bahasa sunda dengan menggunakan metode kawihkaulinan sunda Jurnal PGPAUD Kampus Cibiru 3(3)

[3] O Abdel-Hamid et al 2014 Convolutional neural networks for speech recognition 22(10) 1533-1545

[4] D Palaz MM Doss and R Collobert 2015 Convolutional neural networks-based continuous speech recognition using raw speech signal IEEE International Conference on. IEEE Acoustics, Speech and Signal Processing (ICASSP)

[5] U Nallasamy F Metze and T Schultz 2013 Enhanced polyphone decision tree adaptation for accented speech recognition 1902

[6] L Deng, D Yu and GE Dahl 2015Deep belief network for large vocabulary continuous speech recognition US Patent (8)

[7] A Graves and N Jaitly 2014 Towards end-to-end speech recognition with recurrent neural networks, Proceedings of the 31st International Conference on Machine Learning (ICML-14)

[8] S Bhardwaj et al 2013 Estimation of solar radiation using a combination of Hidden Markov Model and generalized Fuzzy model Solar Energy 93 43-54

[9] H Wu J Wang and X Zhang 2017 Combining hidden Markov model and fuzzy neural network for continuous recognition of complex dynamic gestures The Visual Computer 33(10) 1265-1278

[10] Yulita I N et al 2012 Fuzzy Hidden Markov Models for Indonesian Speech Classification JACIII 16(3) 381-387

[11] Y Ding J Tang X Xu and J Zhang, 2015 Application of linear predictive coding for doppler through-wall radar target tracking IEEE Geoscience and Remote Sensing Letters 12 (6) 1317–1321

[12] L. Grama and C Rusu 2017 Audio signal classification using linear predictive coding and random forests International Conference on IEEE Speech Technology and Human Computer Dialogue (SpeD) 1–9

[13] Yulita I N & Wasito I 2013 gCLUPS: Graph clustering based on pairwise similarity In Information and Communication Technology (ICoICT), International Conference of (pp. 77-81) IEEE

[14] M Gong et al 2008 Fuzzy c-means clustering with local information and kernel metric for image segmentation IEEE Transactions on Image Processing 22(2) 573-584

[15] Massart Thierry Meuter Cedric V B Laurent 2008 On the complexity of partial order trace model checking Inform Process Lett 106 120-126