TELKOMNIKA, Vol.15, No.1, March 2017, pp. 322~327
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v15i1.4656

Modeling Text Independent Speaker Identification with Vector Quantization

Syeiva Nurul Desylvia*[1], Agus Buono[2], Bib Paruhum Silalahi[3]
[1,2]Department of Computer Science, Bogor Agricultural University, Indonesia
[3]Department of Mathematic, Bogor Agricultural University, Indonesia
Corresponding author, email: desylvia_sn@apps.ipb.ac.id*[1], agusbuono@apps.ipb.ac.id[2], bibparuhum@gmail.com[3]

Abstract

Speaker identification is one of the most important technologies nowadays. Many fields such as bioinformatics and security are using speaker identification. Also, almost all electronic devices are using this technology too. Based on number of text, speaker identification divided into text dependent and text independent. On many fields, text independent is mostly used because number of text is unlimited. So, text independent is generally more challenging than text dependent. In this research, speaker identification text independent with Indonesian speaker data was modelled with Vector Quantization (VQ). In this research VQ with K-Means initialization was used. K-Means clustering also was used to initialize mean and Hierarchical Agglomerative Clustering was used to identify K value for VQ. The best VQ accuracy was 59.67% when k was 5. According to the result, Indonesian language could be modelled by VQ. This research can be developed using optimization method for VQ parameters such as Genetic Algorithm or Particle Swarm Optimization.

Keywords*: speaker identification, text independent, vector quantization, Indonesian speaker, K-Means clustering*

1. Introduction

Nowadays, speech based technology has been widely used, such as speech based password at smartphone and PC/laptop to biometric and security applications. Those technologies are example of speaker recognition implementation. One of speech recognition branch is speaker identification which is to identify new speaker based on developed speaker model without known hypothesis that the new speaker is. Then, speaker recognition is divided into text dependent and text independent. Text independent system only depends on vocal tract characteristic from each speaker and there is no assumption about speech context. Text dependent speaker recognition does recognition based on defined words [1]. The possibility of research based on speaker identification is still wide because this technology is still developing rapidly. The example of those research including modification of current method and finding new method.

Research in speaker recognition is one of rapid development research. Reference [2] has been modelled Hidden Markov Model (HMM) combined with Particle Swarm Optimization (PSO) as new approach for speaker recognition system. Neural Network (NN) also has been used for text independent speaker identification [3, 4]. It was Feed Forward Backpropagation (BPNN) combined with Wavelet Entropy as feature extractor. Accuracy of this system was about 90% with Arabic speech data. Wavelet Transform (WT) also used alongside with Linear Prediction Coding (LPC) [5]. At that research, some extraction feature method for speaker identification was compared with proposed method. The highest resulted accuracy was 97.36% at one of proposed method which was Wavelet Packet Transform with Average Framing LPC Feature Extraction (WPLPCF). Comparison between Gaussian Mixture Model (GMM) and Probabilistic Neural Network (PNN) also has been done with feature extraction using Average Framing LPC Feature Extraction (AFLPC). On the other hand, research about text independent speaker verification to reduce feature size also has been done with Genetic Algorithm and Ant Colony Optimization [6].

Received September 25, 2016; Revised December 15, 2016; Accepted January 16, 2017

Template based method such as Vector Quantization (VQ) has been widely developed in many research. For example, [7] modified VQ at similarity search between template training vectors with testing vectors.

According to mentioned research, VQ is capable to be used for speaker recognition system. In this research, VQ will be used to model Indonesian language for text independent speaker identification.

The aim of this research is for modeling Indonesian language data VQ. To get expected result, stages will be done are data collection, preprocessing, dividing training and testing data, feature extraction, modeling data with VQ, testing, and evaluation.

2. Research Method

Figure 1 shows the method of this research. Input model is a speech data and output model is recognized speaker id according to testing data. This input output process showed in Figure 2.

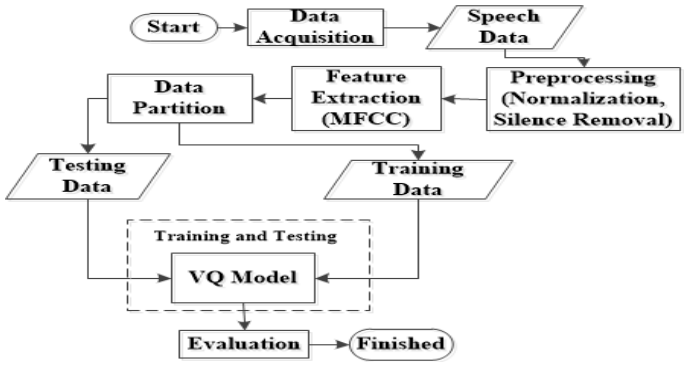

Figure 1. Research Method

2.1. Data Acquisition

Data used in this research are speech data consisting of 985 words. The data recorded using PC head phone at Computational Intelligence laboratory, Computer Science IPB. Speakers are 3 women and 2 men with age range about 25 to 35 years old. The speakers' condition is healthy and there is no cover when recording. Application used is Audacity 2.1.2.

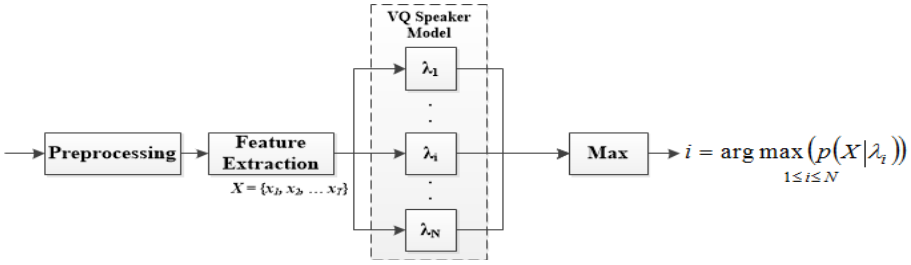

Figure 2. System Input and Output

2.2. Preprocessing

This phase divided into 2 steps which are normalization and silence removal. At normalization, speech data normalized at range -1 to 1. It is to reduce mic effect in recording process. Normalization method used is Min-max Normalization according to this formula [8]:

$$v_i' = \frac{v_i - \min_A}{\max_A - \min_A} (new\_max_A - new\_min_A) + new\_min_A. \quad (1)$$

v_i': New value according to resulted normalization.
v_i: The value to be normalized.
min_A: Minimum value in data.
max_A: Maximum value in data.
new_max_A: New maximum value in data (1).
new_min_A: New minimum value in data (-1).

The objective of silence removal is to get the data that is ready to process without any silence. Silence removal conducted using Audacity 2.1.2.

Table 1. Data Duration Before and After Preprocessing

Speaker	Before Preprocessing (Minutes)	After Preprocessing (Minutes)
1	07.55	05.30
2	08.21	05.39
3	07.09	05.27
4	06.21	05.18
5	08.38	06.20

2.3. Data Partition

In this research, each speech data for each speaker divided into 10 segments. Next, K-Fold Cross Validation (CV) is used to measure model performance based on some sample data partition randomly. In this research, 5-Fold Cross Validation is used.

2.4. Feature Extraction

Prior to feature extraction phase, data integration is carried out. Then, Frequency Cepstral Coefficients (MFCC) is implemented for feature extraction. The number of coefficient is 13 cepstral coefficients.

2.5. Training and Testing

VQ model or centroid model is one of the simplest speaker model text independent. In the application of speech recognition, for example testing data feature vector is symbolized by $X = \{x_1, \dots, x_T\}$ and reference vector symbolized by $R = \{r_1, \dots, v_k\}$ with *D* is average quantization distortion that:

$$D(X,R) = \frac{1}{T}\sum_{t=1}^{T} \min_{1\le k\le K} d(x_t, r_k), \quad (2)$$

$d(.,.)$ is a distance metric such as Euclidean distance $\|x_t - r_k\|$. The smaller the value of equation (2) means highest likelihood between *X* and *R* from the same speaker with *k* is the number of cluster defined and *t* is the number of testing data vector. As a note, $D(X,R) \neq D(R,X)$.

Theoretically, it is possible to implements of all training data vector as a reference vector *R*. But, for computation reason, usually the number of vector is reduced using clustering method such as K-Means. The result of that clustering method is called as codebook (Kinnunen 2010). In this research, the value of *k* at K-Means algorithm is defined by Hierarchical Agglomerative Clustering (Ward).

2.6. Evaluation

In this phase, accuracy calculation is carried out to measure overall system performance. Furthermore, Signal to Noise Ratio (SNR) is used to measure model performance with noisy data. In this research, SNR value applied are 30, 20, 10, 8, 5, and 0.

3. Results and Analysis

3.1. Data Acquisition

Resulted speech data duration and speech vector length from recording are presented in Table 1.

3.2. Pre-processing

The first phase in pre-processing is normalization according to equation (1). The next phase is silence removal using Audacity 2.1.2. The result of this phase is also presented at Table 1.

3.3. Data Partition

The result of 5-Fold cross validation is 40 training data vector and 10 testing data vector for each fold.

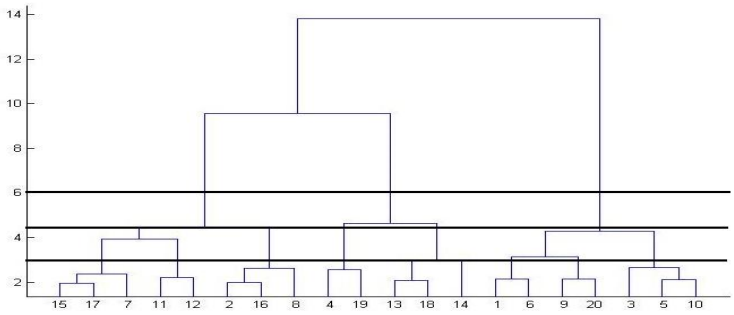

Figure 3. Dendrogram to Define the Value of *k*

3.4. Feature Extraction

The result of this phase is a feature vector matrix which has 13 feature coefficients as defined.

3.5. Training and Testing

In this phase, VQ function is called as much as the number of fold defined. Based on resulted Dendrogram at Figure 3, the number of *k* used are 3, 5, and 9.

Figure 4 presents VQ result without any noise addition to data. In this scenario, VQ gets 100% accuracy. Next, Figure 5 shows VQ result with noise addition at *k* equals to 3. VQ has 100% accuracy at SNR 30 but decreasing at highest SNR. At SNR 20, VQ accuracy is 96% then 40% at SNR 10, 8, and 5. VQ obtains accuracy of 38% at SNR 0. Based on Figure 6 that presents VQ with k equals to 5, VQ recognizes all testing data with accuracy of 100% at SNR 30. At SNR 20, resulted accuracy is 92% then 52% at SNR 10. Accuracy at SNR 8, 5, and 0 are 42%, 40%, and 32%. Figure 7 presents testing result with *k* equals to 9. VQ acquires the best accuracy (100%) at SNR 30 and 20. Then, accuracy decreases at SNR 10, 8, 5, and 0 to be 40%, 40%, 40%, and 20%. Based on resulted accuracy, VQ model is capable to recognize testing data for each speaker well at SNR 30 and 20 but after that, the model performance decreases in conjunction with higher noise ratio.

VQ obtains better average accuracy at *k* equals to 5 which is 59.67% then 59% and 56.67% when *k* equals to 3 and 9.

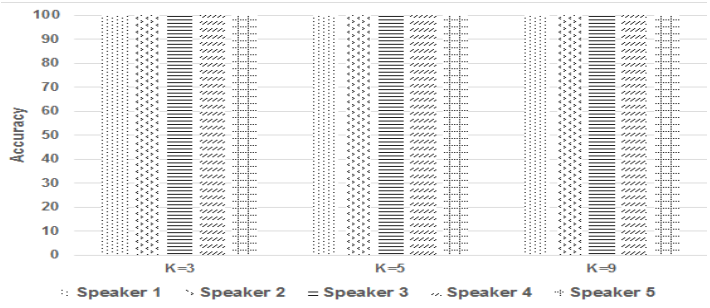

Figure 4. VQ Testing Result without Noisy Data

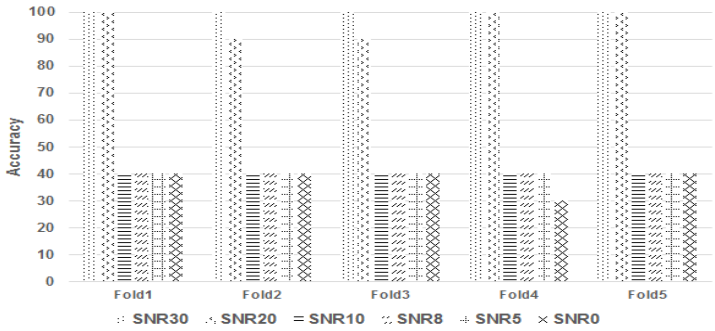

Figure 5. VQ Testing Result with *k* Equals to 3

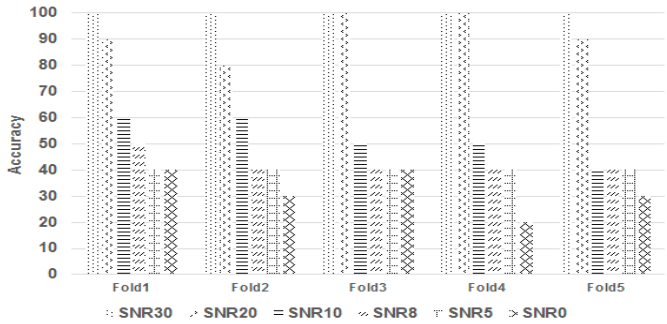

Figure 6. VQ Testing Result with *k* Equals to 5

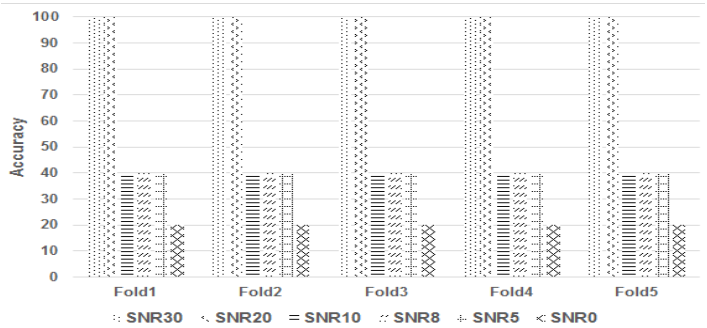

Figure 7. VQ Testing Result with *k* Equals to 9

3.6. Evaluation

Based on testing result, VQ accuracy is 59% at *k* equals to 3. Furthermore, VQ at *k* equals to 5 obtains 59.67% and VQ gets 56.67% at *k* equals to 9. It can be concluded that althought VQ still needs more improvement; VQ can be used to model Indonesian language for text independent speaker identification. Another insight is at Figure 8 which shows VQ's average accuracy based on the value of *k*. It can be observed that VQ accuracy is the best when *k* equals to 5. The resulted accuracy of VQ at *k* equals to 3, 5, and 9 consecutively are 59%, 59.67%, and 56.67%.

VQ still needs improvement because it only uses clustering data mean to model each speaker.

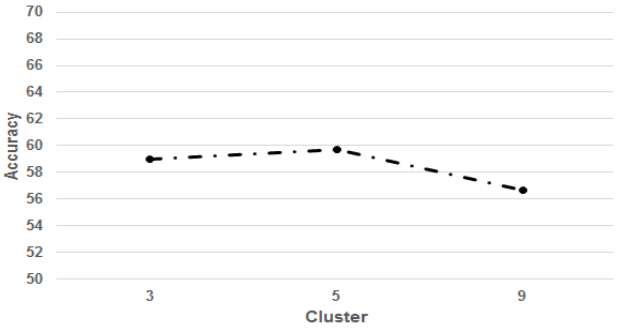

Figure 8. VQ Testing Result based on *k* Values

4. Conclusion

Based on training and testing, Vector Quantization (VQ) best accuracy is 59.67% at speaker identification text identification model in Indonesian Language. This is because VQ only model each speaker only by their clustering mean.

VQ still needs improvement. For future research, the application of optimization methods such as Maximum Likelihood, Genetic Algorithm, or Particle Swarm Optimization can be developed.

References

[1] Beigi H. Fundamentals of Speaker Recognition. New York (US): Springer. 2011: 3-5.
[2] Najkar S, Razzazi F, Sameti H. A novel approach to HMM-based speech recognition systems using particle swarm optimization. *Math Comput Model*. 2010; 52(2010): 1910-1920.
[3] Daqrouq K. Wavelet entropy and neural network for text-independent speaker identification. *Eng Appl Artif Intel*. 2011; 24(2011): 796-802.
[4] Daqrouq K, Tutunji TA. Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers. *Applied Soft Computing*. 2015; 27(2015): 231-239.
[5] Daqrouq K, Al Azzawi K. Average framing linear prediction coding with wavelet transform for text-independent speaker identification system. *Computers and Electrical Engineering*. 2012; 38(2012): 1467-1479.
[6] Nemati S, Basiri ME. Text-independent speaker verification using ant colony optimization-based selected features. *Expert Syst Appl*. 2011; 38(2011): 620-630.
[7] Permana I, Buono A, Silalahi BP. Similarity measurement for speaker identification using frequency of vector pairs. *Telkomnika*. 2014; 12(8): 6205-6210.
[8] Han J, Kamber M, Pei J. Data Mining Concepts and Techniques. 3rd Edition. Amsterdam (NL): Morgan Kaufmann. 2012: 113-114.