Indonesian Journal of Electrical Engineering and Computer Science
Vol. 1, No. 3, March 2016, pp. 590 ~ 596
DOI: 10.11591/ijeecs.v1.i3.pp590-596

A Comparative Study on Similarity Measurement in Noisy Voice Speaker Identification

Inggih Permana
Department of Information System, Faculty of Science and Technology,
State Islamic University of Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia
e-mail: *1inggihpermana@uin-suska.ac.id

Abstract

One of important part on speaker identification is the measurement of sound similarity. This study has compared two of the similarity measurement techniques in the noisy voice.First techniqueis done by using smallest vector sum of pairs and second technique is done by using frequency of occurrence of smallest vector pairs. Noise in the voice can reduce accuracy of speaker identification significantly. To overcome this problem, the two of similarity measurement was combined with Least Mean Square (LMS) for remove noise. Results of the experiments showed that the use of LMS can improve the accuracy of speaker identification at the two of similarity measurement techniques. Second technique produces better accuracy than first technique. Experimental result also showed improvement of LMS learning rate can improve the accuracy of speaker identification.

Keywords*: LMS, noisy voice, sound similarity measurement, speaker identification*

1. Introduction

Speaker recognition is part of the sound processing that aims to find out who is speaking. Speaker recognition is divided into two parts, the speaker identification and speaker verification. Speaker identification is a manner to identify someone from the existing voice, whereas speaker verification is a manner to verify a claim against an identity through certain words [1]. This study focuses on speaker identification.

One of important part on speaker identification is the measurement of sound similarity. In this part will be determined owner of the voice that identified. In the previous study [2] has made modifications to the sound similarity measurement techniqueby selecting the codebook that has the most of smallest distance with input vectors to produce a better identification accuracy. But the technique is not resistant to noisy sound. This study aims to improve the capability of that technique by adding active noise canceling (ANC) in the pre-process of data.Research conducted by Permana *et al.* [3] showed the ANC can improve the accuracy of speaker identification. The ANC method used in this research is least means square (LMS).

This study used a mel frequency cepstral coefficient (MFCC) as a feature extraction and self-organizing map (SOM) as a codebook maker. MFCC chosen because of the way it works is based on the frequency difference can be captured by the human ear so that it can represent how people receive sound signals [4]. MFCC is often used because it is considered a better performance than other methods, such as in terms of reduced error rates. SOM chosen because it has been successfully applied to high-dimensional data [5]. This is the reason for using SOM, because the results of the MFCC vectors can be high dimension.

2. Research Method

Broadly speaking, this study is divided into three parts. The first part is making of codebook using training voice data that are not given noise. The second part is a measurement of similarity to the codebook that has been made.Voice data used in this part is the test voice data that has been given the noise. This study used white noise with various values below 6.5 dB. At this part LMS is used as a data preprocessing to remove noise. There are two similarity measurement techniques used in this study, that are by using smallest vector sum of pairs [6, 7] (the first technique) and by using frequency of occurrence of smallest vector pairs [2] (the

Received December 3, 2015; Revised January 24, 2016; Accepted February 9, 2016

second technique). The last part is the comparison and analysis of the similarity measurement techniques used. For more details, see Figure 1.

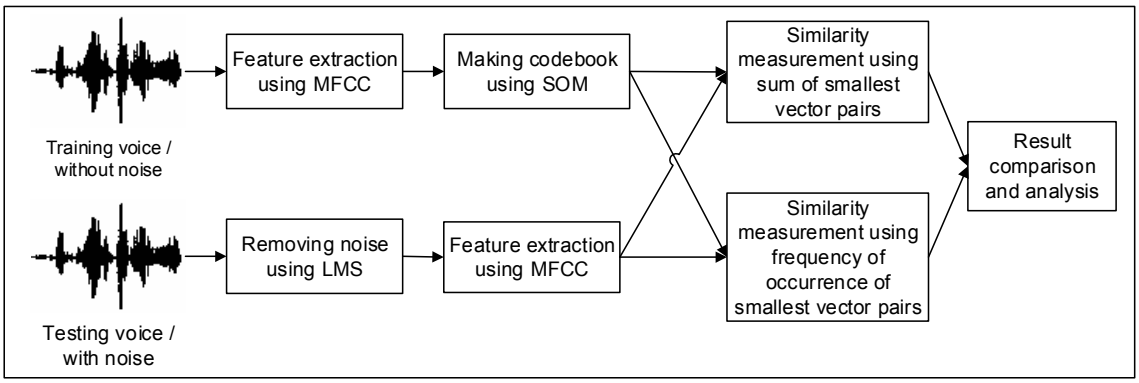


Figure 1. Illustration of research method

Experiments in this study will be performed in several combinations of parameters. Learning rate of ANC that is attempted 0.1, 0.3, 0.5, 0.7 and 0.9. MFCC coefficients that is attempted 13, 15 and 20. SOM cluster number that is attempted is 9, 16, 25, 36, 49, 64, 81 and 100.MFCC frame length is 12.5 ms. Overlap of MFCC is 0.4. Topology of SOM is hexagonal. SOM iteration number is 1000.

At each combination of parameters one voice files that owned by each speaker will be used to create the codebook. After that, testing performed using all voice data. All voice data used has been removed silent time. This is done 5 times so that all voice files for each speaker ever be the data to create the codebook. For each experiment are calculated the resulting accuracy. After all the experiments carried out in a combination of particular parameters, then computed the average of accuracy. This accuracy is used as the level of speaker identification ability.

2.2. Voice Data

Voice data used is ever used by Reda [8] in their study. The voice data consists of 83 speakers, which are divided into 35 female speakers and 48 male speakers. The speakers are Indian citizens of different backgrounds. Each speaker has 5 voice files in wav format. The voice file length is 1 to 39 seconds. The words that speak by the speaker is a random combination of numbers. Recording is done on the phone using an IVR system (Interactive Voice Response). Sampling rate used is 8000 Hz.

2.3. Similarity Measurement

This study uses two measurement techniques similarity. In the first techniques [6, 7], each input vector is measured the distance with vectors that exist in a particular speaker codebook. Choose a pair of vectors which has the smallest distance for each input vector. Sum all the minimal pairs that obtained. Perform these processes for all existing speaker codebook. After that, choose the codebook with the most minimal sum as speakers representing the voice identified. Illustration of first techniques can be seen in Figure 2.

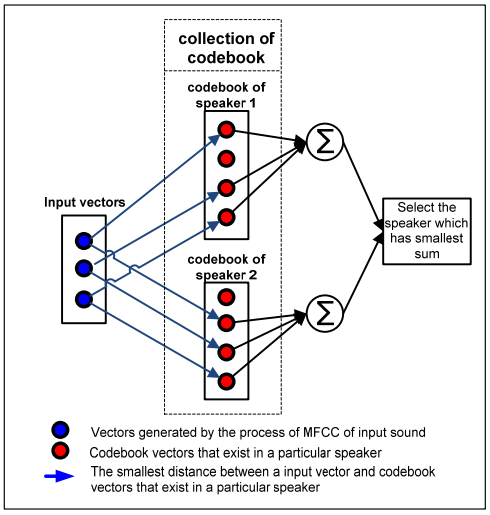


Figure 2. Previous similarity measurement techniques [2]

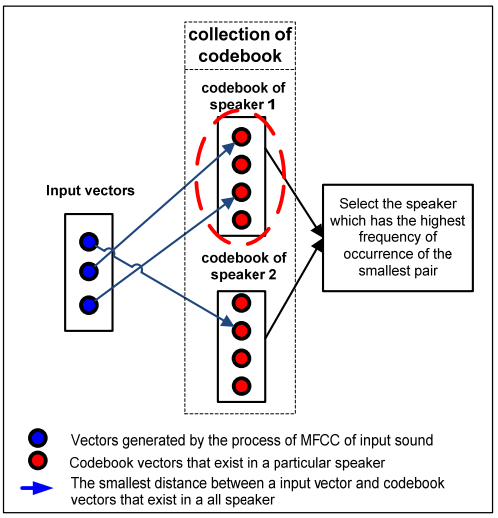


Figure 3. Proposed similarity measurement techniques [2]

In thesecond techniques [2], the input vectors are not only measured the distance to the particular speaker codebook, but it will be measured with all vectors that exist in all available speaker codebook. The smallest distance selected from the input vector to one of a collection of vectors that exist in the available codebook. Codebook vector which causes the smallest distance will be selected as the pair of the input vector. After that, select the codebook that has the highest frequency pair as speakers representing the input voice. Illustration of the second techniques can be seen in Figure 3.

2.4. Mel Frequency Cepstral Coefficient (MFCC)

This research used a type MFCC-FB40 [9] because it has the equal error rate (EER) and decision cost function (DCFopt) is lower than the other types of MFCC [10]. Illustration MFCC stages can be seen in Figure 4.

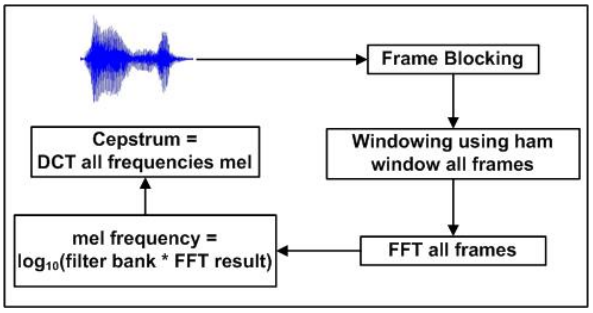


Figure 4. Illustration of the MFCC process [2]

The first step in the MFCC process is divide the incoming signal into multiple frames. The second step is the smoothing of each frame to minimize non-continuous signal using hamming window. The third step is to convert the voice signal from the time domain to the frequency domain using the fast fourier transform (FFT). The fourth step is to change the frequency of the FFT results into mel scale. The final step is to restore the signal from the time domain to the frequency domain using the discrete cosine transform (DCT).

2.5. Self Organizing Map (SOM)

SOM or also known as Kohonen is one type of artificial neural network (ANN) with unsupervised learning system. SOM is very effective to create an internal representation of space that is organized for the various features of the input signal [11]. SOM assumes topology structure among clusters of units, it is run by a human brain but is absent in some other ANN [12].

The first step of training process using SOM is determine the number of clusters to be generated. After that, the next step is to create a vector for each cluster. Vectors cluster are given initial weight. Find the smallest distance between the input vectors and the cluster vectors. Cluster vector that causes the smallest distance is the winner vector. Update the weight vector of the winner using Equation 1.

$$w_{ij}(new) = w_{ij}(old) + \alpha[x_i - w_{ij}(old)] \quad (1)$$

Where w is the weight of the unit in the output layer, x is the input data and α is the learning rate.

2.6. Least Mean Square (LMS)

ANC method used in this study is the LMS (Least Mean Square). LMS is applying the gradient descent. This method was first proposed by Widrow and Hoff [13]. Steepest descent, which is one method that implements gradient descent actually been very good to generate optimal weights, but this method requires a true gradient at each step. LMS can overcome these shortcomings because LMS can instantly estimate the gradient at each step.

The first step of LMS are create a filter and initialization the weight (w) ofthe filter. After that specify a value of the learning rate (α). Calculate the anti-noise using the Equation 2. Then, calculate the residual signal using Equation 3. The last step is change the weights using equation 4.

$$y_i = \sum_{j=1}^{M} w_j u_j \quad (2)$$

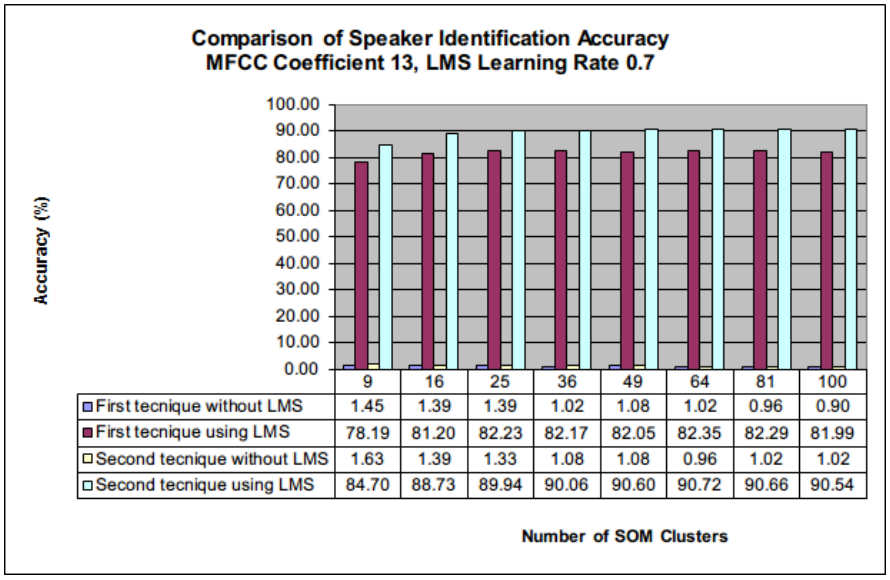
$$e_i = d_i - y_i \quad (3)$$

$$w_j(new) = w_j + 2\alpha e_i u_j \quad (4)$$

Where y is anti-noise, u is reference noise, d is incoming voice signal, and e is residual signal.

4. Results and Analysis

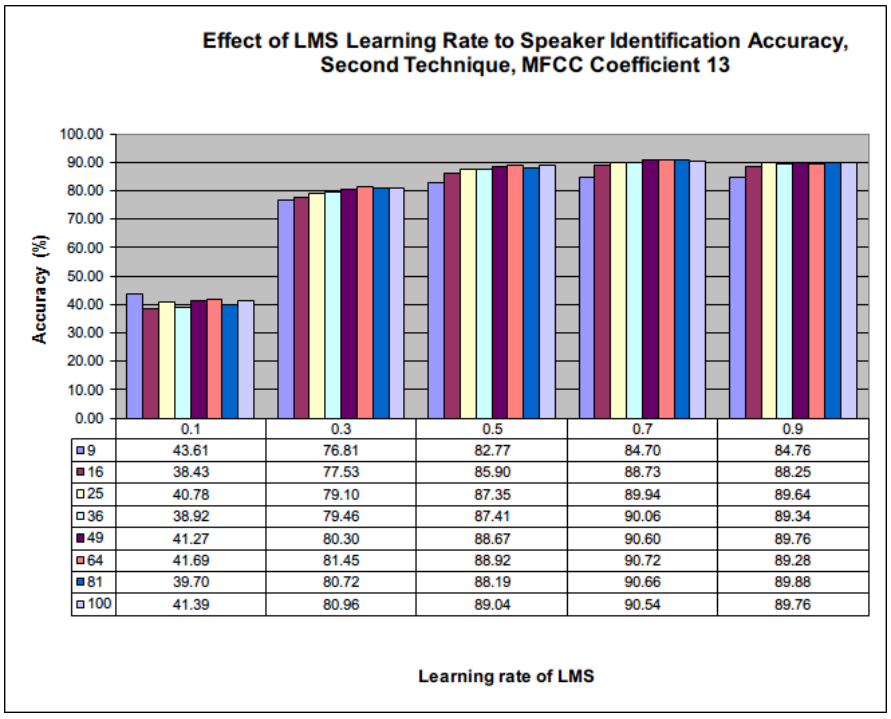
The graph in Figure 5 shows the highest accuracy in the speaker identification for noisy test data that do not use the LMS in data preprocessing is very low, 1.45% in the first similarity measurement technique and 1.63% in the second similarity measurement technique. Both accuracy occurred in the number of SOM clusters is 9.



	9	16	25	36	49	64	81	100
First tecnique without LMS	1.45	1.39	1.39	1.02	1.08	1.02	0.96	0.90
First tecnique using LMS	78.19	81.20	82.23	82.17	82.05	82.35	82.29	81.99
Second tecnique without LMS	1.63	1.39	1.33	1.08	1.08	0.96	1.02	1.02
Second tecnique using LMS	84.70	88.73	89.94	90.06	90.60	90.72	90.66	90.54

Figure 5. Comparison of Speaker Identification Accuracy

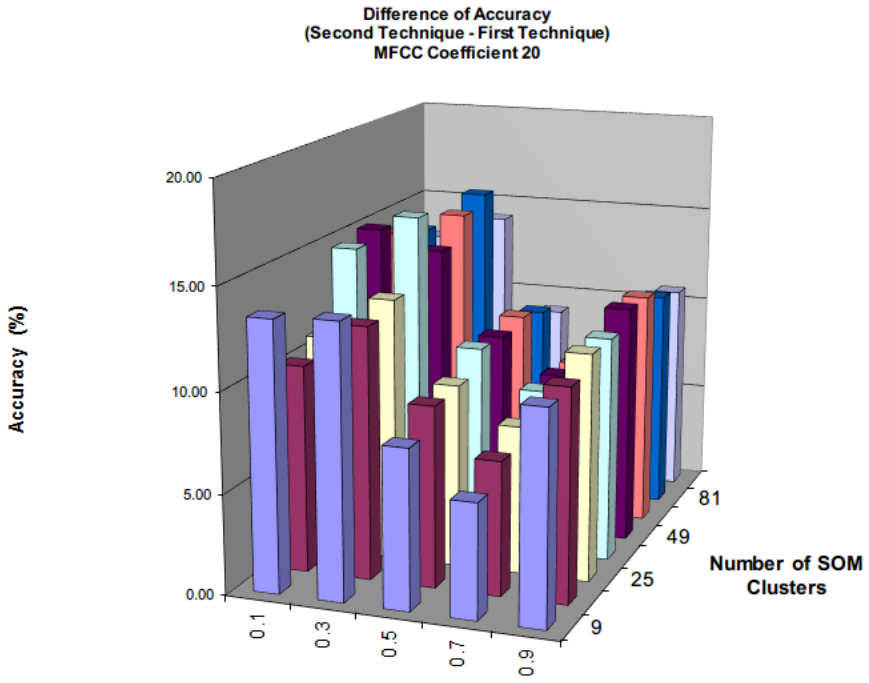
The graph in Figure 5 shows the speaker identification accuracy becomes greatly increased after the addition of LMS algorithm on the data preprocessing. The highest accuracy in the first technique is 82.35%. The highest accuracy in the second technique is 90.72%. Both of highest accuracy occurs at SOM with the number of clusters is 64.Both of highest accuracy also showed an increase in accuracy between the first technique and the second technique in which use LMS on preprocessing of data is quite significant with the highest accuracy improvement 8.37%.



	0.1	0.3	0.5	0.7	0.9
9	43.61	76.81	82.77	84.70	84.76
16	38.43	77.53	85.90	88.73	88.25
25	40.78	79.10	87.35	89.94	89.64
36	38.92	79.46	87.41	90.06	89.34
49	41.27	80.30	88.67	90.60	89.76
64	41.69	81.45	88.92	90.72	89.28
81	39.70	80.72	88.19	90.66	89.88
100	41.39	80.96	89.04	90.54	89.76

Figure 6. Effect of LMS Learning Rate

Figure 6 shows the effect of LMS learning rate on second technique. When the value of learning rate is 0.1, the resulting accuracy is very low which the highest accuracy is only 47.89%. When learning rate increased to 0.3 and above, the accuracy increased very significantly. The highest accuracy occurs when the learning rate increased to 0.7, which the accuracy is 92.47%.



	0.1	0.3	0.5	0.7	0.9
9	13.43	13.61	7.95	5.72	10.54
16	10.36	12.59	9.04	6.69	10.60
25	10.96	13.13	9.16	7.41	11.33
36	14.76	16.57	10.18	8.31	11.27
49	15.12	14.22	10.00	8.25	11.99
64	14.28	15.48	10.30	8.07	11.87
81	13.86	16.02	9.82	7.17	11.14
100	12.89	14.10	9.10	7.53	10.72

Figure 7. Difference of Accuracy (Second Technique – First Technique)

The graph in Figure 6 shows the difference in the accuracy of the speaker identification between similarity measurement techniques that used. The second technique always produces higher identification accuracy. Difference of highest accuracy found on the learning rate is 0.3 and the number of clusters SOM is 36, which is 16.57%.

5. Conclusion

Based on the results of experiments on noisy voice, the use of LMS can improve the accuracy of speaker identification atcombination of the smallest vector sum of pairs techniques [6, 7] with LMS andthe combination of the frequency of occurrence of smallest vector pairs techniques [2] with LMS. Second combination produces better accuracy than first combination. Improvement of LMS learning ratecan improve the accuracy of speaker identification for all combinations. Experiments in this study showed the best learning rate is 0.7.

References

[1] Togneri R, Pullella D. An Overview of Speaker Identification: Accuracy and Robustness Issues. *Circuits and Systems Magazine*, *IEEE*. 2011; 11(2): 23-61.
[2] Permana I, Buono A, Silalahi BP. Similarity Measurement for Speaker Identification Using Frequency of Vector Pairs. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2014; 12(8): 6205-6210.
[3] Permana I, Buono A, Silalahi BP. *Noise Cancelling for Robust Speaker Identification Using Least Mean Square*. Proceeding. Proceedings of the 1st International Conference on Science and Technology for Sustainability. IcosTechs. 2014; 1: 247-252.
[4] Muda L, Begam M, Elamvazuthi I. Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *Journal of Computing*. 2010; 2(3): 138-143.
[5] Yan J, Zhu Y, He H, Sun Y. Multi-Contingency Cascading Analysis of Smart Grid Based on Self-Organizing Map. *IEEE Transactions on Information Forensics and Security*. 2013; 8(4): 646-6.
[6] Fruandta A, Buono A. *Identifikasi Campuran Nada pada Suara Piano Menggunakan Codebook*. Seminar Nasional Aplikasi Teknologi Informasi. Yogyakarta. 2011; 8-13.
[7] Wisnudisastra E, Buono A. Pengenalan Chord pada Alat Musik Gitar Menggunakan CodeBook dengan Teknik Ekstraksi Ciri MFCC. *Jurnal Ilmiah Ilmu Komputer*. 2010; 14(1): 16-21.
[8] Reda A, Panjwani S, Cutrell E. *Hyke: A Low-Cost Remote Attendance Tracking System for Developing Regions*. Proceedings of the 5th ACM workshop on Networked systems for developing regions. ACM. 2011; 15-20.
[9] Slaney M. *Auditory Toolbox*. Interval Research Corporation, Tech Rep. 1998.
[10] Ganchev T, Fakotakis N, Kokkinakis G. *Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task*. Proceedings of the SPECOM. 2005; 1: 191-194.
[11] Kohonen T. *The Self-Organizing Map*. Proceedings of the IEEE. 1990; 78(9): 1464-1480.
[12] Basheer IA, Hajmeer M. Artificial Neural Networks: Fundamentals, Computing, Design, and Application. *Journal of Microbiological Methods*. 2000; 43(1): 3-31.
[13] Kinnunen T, Li H. An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. *Speech Communication*. 2010; 52(1): 12-40.
[14] Furui S. An Overview of Speaker Recognition Technology. *Automatic speech and speaker recognition*. Springer US. 1996; 31-56.
[15] Alam MJ, Kenny P, Ouellet P, O'Shaughnessy D. Multitaper MFCC and PLP Features for Speaker Verification Using i-Vectors.*Speech Communication*. 2013; 55: 237–251.
[16] Chen SH, Luo YR. *Speaker Verification Using MFCC and Support Vector Machine*. Proceedings of the International Multi Conference of Engineers and Computer Scientists. Hong Kong. 2009; 1: 18-20.
[17] Nakagawa S, Wang L, Ohtsuka S. Speaker Identification and Verification by Combining MFCC and Phase Information. *IEEE Transactions on Audio, Speech, and Language Processing*. 2012; 20(4): 1085-1095.
[18] Davis S, Mermelstein P. *C*omparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*. 1980; 28(4): 357-366.
[19] Steve Y, Odel J, Ollason D, Valtchev V, Woodland P. *The HTK Book, version 2.1*. Cambridge University. 1997.
[20] Skowronski MD, Harris JG. Exploiting Independent Filter Bandwidth of Human Factor Cepstral Coefficients in Automatic Speech Recognition. *The Journal of the Acoustical Society of America*. 2004; 116: 1774–1780.
[21] Kohonen T. Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*. 1982; 43(1): 59-69.