Speaker Recognition Implementation for Authentication Using Filtered MFCC – VQ and a Thresholding Method

Reza Aulia Sadewa[1], Tokorda Agung Budi Wirayuda[2], Siti Sa'adah[3]
[1, 2, 3]School of Computing
Telkom University
Bandung, Indonesia
[1]rezaauliasadewa@gmail.com, [2]cokagung@telkomuniversity.ac.id, [3]sitisaadah@telkomuniversity.ac.id

***Abstract*—This paper explains about authentication mechanism using one of the unique biometric component, the human voice. First of all, the characteristic of the voice is extracted using MFCC then represented by *cepstrum* coefficients. Later, those features forms a model by the VQ method. These methods are modified with a proposed thresholding method to reject the unknown voice and a Butterworth Filter to handle the noise. For the experiment, we used both synthetic and real human voice, or biometric data. Both synthetic and biometric data consist of 10 speaker. Half of the speaker is separated as the unregistered or the untrained voices. Overall, the result shows that the methods is adequate enough to perform a security mechanism. MFCC and VQ combination can truly 100% distinguish the speakers in a closed sample which includes only the registered speaker. Compared to the noise-added data, the noise-filtered data can increase the true acceptance accuracy with a specific filter parameters. The proposed thresholding method is effective enough to reject the unknown voice with approximately 90% true rejection but produces only around 70% true acceptance. Hence, the value of the threshold tolerance, which is to increase the authentication accuracy for the registered speaker, needs to be treated more for the next experiment to find the balance between the acceptance, and the rejection accuracy.**

***Keywords*–MFCC, Butterworth, VQ, LBG, threshold**

I. INTRODUCTION

The human voice is a unique biometric component. One of the research has proven it statistically [9]. The main problem of speaker recognition is how the system can distinguish the voices between speakers and avoid access to the untrained or unknown voices. The existence of the noise in a voice signal has a potential to make the extracted features of the voice is rather unrepresentative.

There has been similar researches [4, 6], but it seems that they don't implement the method to reject the untrained speaker, so the result will not be valid if the untrained voice is included to their experiment. The untrained voice will still be treated as the most similar voice which has been trained before. This is a problem if the system is applied as a security mechanism.

The objective of this work is to build a system which overcome those problems. The other objective of this work is to find out whether the same speaker affect the MFCC features regardless of the words spoken.

This paper is organized as follows. Section II explains the key resume of the algorithms and methods which are used in the system. Section III describes the overall process of the system as a flowchart diagram. Section IV is the summary of the experimental results which is described in several points. The last, Section V, provide the conclusion and key weakness of the system to facilitate the future research.

II. LITERATURE REVIEW AND STUDY

A. Speaker Recognition

Basically a speaker recognition consists of two procedures, training and verification phase. In the training phase a model which contain the characteristic or features and threshold value of each speaker is gained and stored. When the verification phase occured, the input person voice will be compared to all of the stored models and the system will choose one model with the most similar characteristic. The final step, if the characteristic value is in the range of the threshold, the person is authenticated, otherwise the person is treated as an unknown person.

B. Butterworth Filter

The characteristic of the Butterworth filter is that this filter can flatten the signal in the given cut-off frequency [7]. The type of filter for this work is a low-pass filter, so the filter starts to work when the signal reach above the cut-off frequency [8]. This filter is applied for every ampitudes value in the time domain signal [1]. The equation is shown in (1)

$$Y(n) = a_0\,X(n) + a_1\,X(n-1) + a_2\,X(n-2) - b_1\,Y(n-1) - b_2\,Y(n-2) \quad (1)$$

where $Y(n)$ is the n^{th} amplitude value after the filter is applied to $X(n)$ with the use of a and b as the filter coefficient. The a dan b coeficients are obtained from the real

and imaginary value of the Butterworth filter, which is shown in (2).

$$u_m = \frac{1 - x_m^2 - y_m^2}{(1 - x_m^2)^2 + y_m^2} \text{ and} \quad (2)$$
$$v_m = \frac{2\, y_m}{(1 - x_m)^2 + y_m^2}$$

where x_m dan y_m is in equation (3)

$$x_m = \tan\frac{\omega_c}{2}\cos\frac{2m+1}{2N}\pi \text{ dan } y_m \quad (3)$$
$$= \tan\frac{\omega_c}{2}\sin\frac{2m+1}{2N}\pi$$

where $= 0, 1, 2, .., 2N - 1$; N is order of the filter , ω_c is the frequency. The amount of order affect the filter sensitivity against the cut-off frequency. The more order, the better the filter cleans the noise.

C. Mel Frequency Cepstral Coefficient (MFCC)

MFCC is a feature extraction method that represent the characteristic of the signal as coefficients. These coefficient is obtained by applying a number of triangular-shape mel filter into the signal to get the amount energy from certain frequency range in the mel scale, shown in Fig. 1. This method mimic the perception of the human hearings by reshaping the mel filter to logaritmic form above 1000Hz and linear form below 1000Hz [4].

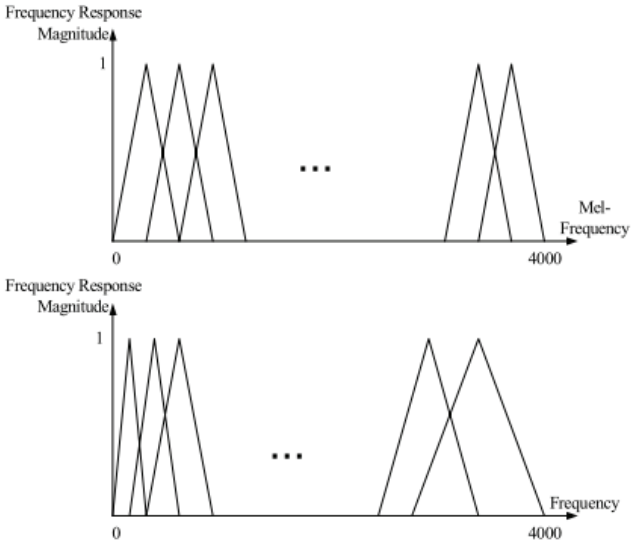

Fig. 1. Mel frequency filter in normal and mel scale [3]

Below are steps of MFCC :

1. *Pre – emphasis*

This step will flatten or stabilize the signal [5] with the pre-determined pre – emphasis constant. The equation is shown in (4)

$$s_n' = s_n - a s_{n-1} \quad (4)$$

where s_n is the n^{th} sample or amplitude and a is the pre – emphasis constant. The effect of the pre – emphasis is visualized in Fig. 2.

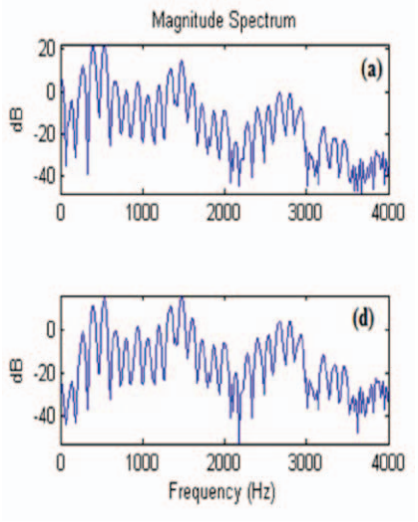

Fig.2. Pre – emphasis result (d) of the signal (a) [5]

2. *Framing*

The voice signal then divided into frames . The length of each frame must short enough, around 20 to 40 ms, to ensure the signal stationarity of every frame [6]. Beside, there is an overlap, half of the frame length, between adjacent frames to make stability between frames [3].

The further steps of MFCC are applied for every frame.

3. *Windowing*

Due to the effect of the framing process, there is a signal discontinuity at the sides of every frame [3]. A window function is applied to refine all of the frame sides. The equation of the Hamming Window is shown in (5).

$$Ham(n) = 0{,}54 - 0{,}46\ \cos\left(2\pi\frac{n}{N-1}\right) \quad (5)$$

where N is the number of samples in every *frame* and $n = 0, 1, 2, .., (N - 1)$.

4. *Fast Fourier Transform (FFT)*

FFT is a fast algorithm to compute the Discrete Fourier Transform (DFT) [2]. The DFT is shown in equation (6).

$$F_k \equiv \sum_{j=0}^{N-1} f_j \, e^{-\frac{2\pi i k(j)}{N}} \quad (6)$$

where F_k is the resulting coefficient from k^{th} DFT , f_j is j^{th} sample from the time domain signal, N is total sample, $j = k$, and i is a complex number. The complex number consists of the real part, which represents the magnitude, and the imaginary part, which represents the phase.

Overall, this process transform the time - domain signal to the frequency - domain signal. The maximum frequency range is half of the sampling rate. The DFT is required for the the mel filter application in the next step.

5. *Mel frequency filter bank*

The signal then converted to the mel – spaced frequency , as shown in equation (7), for the application of the triangular mel filter.

$$Mel(f) = 2595 \, \ln\left(1 + \frac{f}{700}\right) \quad (7)$$

where $Mel(f)$ is the mel frequency value of frequency f.

The result of this step is mel filter banks which represent the energy value for every frequency range.

6. *Non linear transformation*

The purpose of this step is to simulate the human hearing which is more critical in the lower frequency than in the higher frequency, by taking the the mel filter banks into the logaritmic spacing, as shown in equation (8).

$$f'_k = \ln f_k \quad (8)$$

Where f is the mel frequency filter bank, $= 1,2,..,K$; and K is the number of mel frequency filter bank in a frame.

7. Discrete cosine transform (DCT)

DCT transform the signal back to the time domain. The result is a number of cepstrum coefficients for each frame. The DCT equation is shown in (9)

$$c_n = \sum_{k=1}^{K} (f'_k) \cos\left[n(k - 0{,}5)\frac{\pi}{K}\right] \quad (9)$$

where K is the number of mel frequency filter bank, f'_k is obtained from equation (9), $n = 1, 2, .., N$, so that N number of cepstrum coefficients, known as the feature vectors, are derived from each frame. The first coefficient can be ignored because of its irrelevancy [10].

D. Vector Quantization (VQ)

The main purpose of VQ is to reduce big amount of vector data by representing them with their centroid only to speed up the verification process [4]. The centroid's coordinates of a cluster is obtained by taking the average of all vector's coordinate in the cluster.

A group of cluster is called a codebook, which is a characteristic representation of a single person's collection of voice. The process of generating centroids of the codebook is done by using LBG [Linde, Buzo, and Gray, 1980] algorithm. A visualization of a codebook is depicted in Fig. 3.

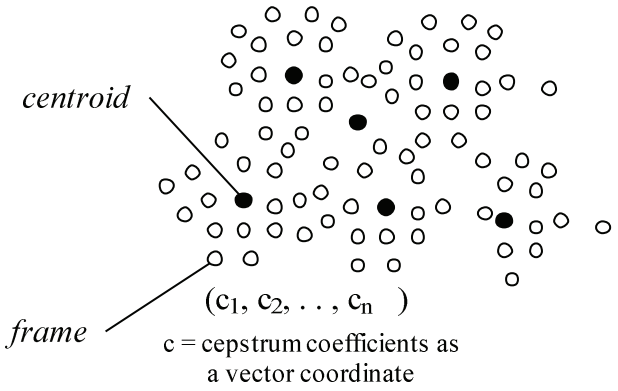

Fig. 3. A single codebook which contain feature vectors and centroids

A parameter which represents the similarity between the input feature vectors and a codebook is a distortion measure. The smaller the distortion, the higher the similarity is. The distortion is shown in equation (10)

$$D_Q(I, V) = \sum_{t=1}^{T} \min d(i_t, v_k) \, ; 1 \leq k \leq K \quad (10)$$

where D_Q is the distortion of the input feature vectors I ,which consist of T vektors, to the *codebook* V , which consist of K *centroids*, i_t is the t^{th} input vector and d is an Euclidean distance which is shown in equation (11).

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \quad (11)$$

where p and q are vectors which their coordinates are $p_{1..n}$ and $q_{1..n}$. In other words, the distortion between a codebook and a group of feature vectors is a sum of the Euclidean distances between each feature vector to the nearest centroid in a codebook. This procedure is visualized in Fig. 4.

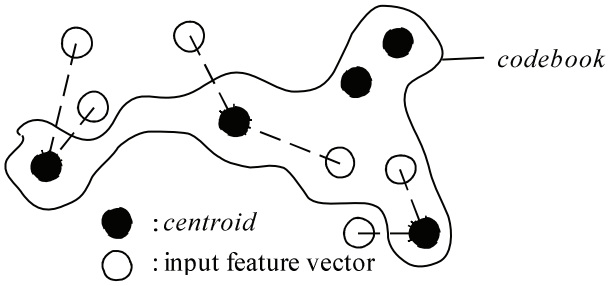

Fig. 4. Distortion between input feature vectors and a codebook

where the distortion value is the sum of the distances which is depicted as the dashed lines.

The steps of the LBG algorithm [4] is :

1. Generate a single centroid from all of the feature vectors
2. Multiply the number of centroid by 2 using the rule shown in (12)

$$y_n^+ = y_n(1 + 0.01)$$
$$y_n^- = y_n(1 - 0.01) \quad (12)$$

where y_n is n number of centroid in a certain iteration.

3. Each vector is assigned to its nearest centroid
4. Repeat step 3 and 4 until the centroids satisfy the condition shown in (13)

$$\sum_{x=1}^{X} D'_{Q_x} - \sum_{x=1}^{X} D_{Q_x} > 0{,}1 \quad (13)$$

where D'_{Q_x} is a distortion of a *cluster* after the 4th step whereas D_{Q_x} is the distortion before the 4th step and X is the number of the cluster in the codebook.

5. Back to the 2nd step and the next procedures are repeated until the determined number of centroid is reached.

Because of the multiplication of 2 in every iteration, the LBG algorithm always produces number of centroids with the power of 2.

E. Thresholding Method

By using just the VQ method, a person's voice whose codebook has not been generated in the training phase before, will also be authenticated as a voice that has the lowest distortion to a certain codebook. On the right perspective, the voice should be rejected. Therefore, we proposed a thresholding method to overcome this problem.

A threshold of each codebook is trained after the codebook of certain person or speaker has been generated. The threshold consist of minimum and maximum distortion value which is obtained by authenticating several voices of the same speaker. In the verification phase, the resulting lowest distortion is validated using rule (14).

$$(a \leq D_Q \leq b) \quad (14)$$
$$\rightarrow \textit{the voice has been registered and valid}$$

where a and b is the threshold value whereas D_Q is the distortion between the input voice to the most similar *codebook*. As an optional procedure, the range of the threshold is increased by adding the difference between the maximum and the minimum of the threshold value that is divided by a chosen constant number as the divisor.

III. SYSTEM'S FLOW

Overall, the system in this work described in Fig. 5.

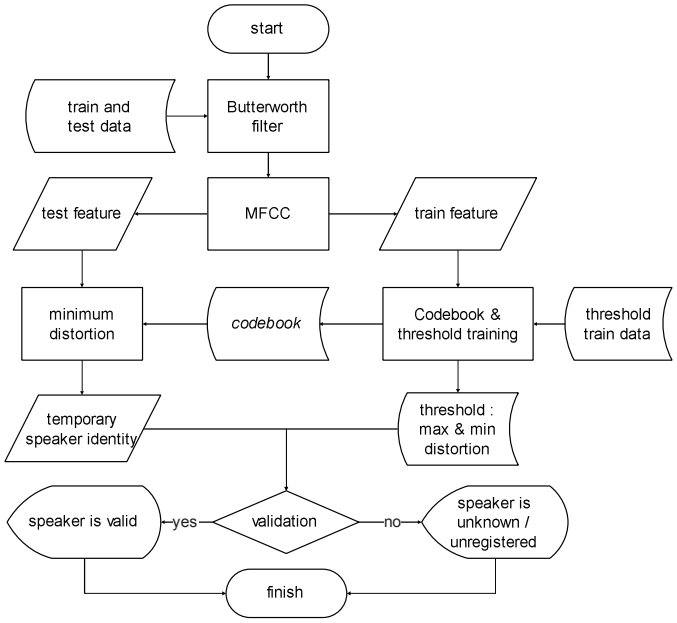

Fig. 5. System's flow

Both the train and test data is feature-extracted and noise-filtered. The difference is that the train data is formed into a codebook and threshold-trained first before it is being compared to the test data later on in the verification phase.

IV. EXPERIMENT RESULTS

The experiment is divided into two types of dataset : synthetic and biometric voices dat.a with 10 speakers each. The number speaker is halfly separated as the untrained voice. The synthetic data is made using the text – to – speech

provided at www.acapela-group.com. The pitch, formant, and the volume are varied using Cockos Reaper software to generate a lot of data. The biometric data is a real human voice which is recorded using a microphone with quality of 44,1 kHz.

The summary of the experiment results, both the synthetic and biometric dataset, are :

1. Just to distinguish the speakers, MFCC and VQ can fully authenticate every test data as the actual speaker. However, without the thresholding method, all of the untrained data are falsely authenticated.

2. The proposed thresholding method can truly reject around 95% of the untrained speaker, but parts of the registered speaker are poorly authenticated, around 74%. This can be overcomed by adding the threshold tolerance with the divisor = 4 which increase true acceptance of the trained data up to around 86% but decrease the untrained data true rejection by around 3%.

3. The alteration of the parameter such as number of centroid in a codebook, mel filter, and cepstrum coefficient does affect the recognition result, but it's not very significant. The optimal values are 256 centroids, 23 mel filters, and 12 cepstrum coefficient.

4. One speaker experiment with the addition of the silence and noise to the synthetic voice makes the system reject 100% of the input data.

5. The lowest SNR difference between the original and the filtered signal can be achieved by applying the Butterworth filter to both data. In other words, both train and the test data must be filtered.

6. The experiment result with the noise-added synthetic registered test data which means 5 speaker, a Butterworth filter with order = 10 and cut-off = 1000Hz shows the best result, around 73% true acceptance.

For the experiment with the biometric data it's not recomended to apply the Butterworth filter because both the train and the test data have noise.

V. CONCLUSION AND FUTURE WORK

Based on the experiments, overall, the combination of the methods is adequate enough to identify and distinguish voice of the speakers.

The accuracy of the system can be diminished by the existence of the noise or the amount of silence. The extracted feature of MFCC is depended on both speaker and the word spoken. So a word difference between the trained voice and the verification can reduce the system's accuracy.

The increasing number of centroid in a codebook, cepstrum coefficient, and the amount of mel filter can increase the accuracy of the system, however it's not very significant.

Noise can be removed using the Butterworth filter, but it's parameters has to be appropriate to the conditions of every data, because a wrongly tuned parameter can ruin the characteristic of the MFCC coefficient which leads to an error authentication.

Our proposed thresholding method is enough to reject all of the untrained or unknown data however, it is not very accurate to authenticate the trained data. The addition of the threshold tolerance can increase the authentication for the trained data, but also decrease the rejection of the unknown data.

This field of research, the thresholding method specifically, has to be continued. The balance of the threshold tolerance is not yet been discovered. This means that the optimal determination of the divisor has to be treated in another research. Beside, the number of speaker has to be increased to ensure the effectiveness of this work's methods.

REFERENCES

[1] Alem, N., & Perry, M. (1995). *Design of Digital Low-pass Filters for Time-Domain Recursive Filtering of Impact Acceleration Signals.* Alabama.

[2] Fast Fourier Transform. (1992). In *NUMERICAL RECIPES IN FORTRAN 77: THE ART OF SCIENTIFIC COMPUTING* (pp. 490-502). Cambridge University Press.

[3] HAN, W., CHAN, C.-F., CHOY, C.-S., & PUN, K.-P. (2006). An Efficient MFCC Extraction Method in Speech Recognition. *IEEE*, 145-148.

[4] Kamale, H. E., & Kawitkar, R. S. (2008). Vector Quantization Approach for Speaker Recognition. *International Journal of Computer Technology and Electronics Engineering* , 110-114.

[5] Loweimi, E., Ahadi, S. M., Drugman, T., & Loveymi, S. (n.d.). On the importance of Pre-emphasis and Window Shape in Phase-based Speech Recognition.

[6] Mayrhofer, R., & Kaiser, T. (n.d.). Towards usable authentication on mobile phones: An evaluation of speaker and face recognition on off-the-shelf handsets.

[7] Parashar, A., & Ghosh, P. K. (n.d.). Speech Enhancement and Denoising Using Digital Filters. *International Journal of Engineering and Science Technology*, 1094-1103.

[8] Robertson, D. G., & Dowling, J. J. (2003). Design and responses of Butterworth and critically damped digital filters. *Journal of Electromyography*, 569-573.

[9] Trilok, N. P., Cha, S.-H., & Tappert, C. C. (2004). Establishing the Uniqueness of the Human Voice for Security Applications. *Proceedings of Student/Faculty Research Day, CSIS, Pace University*, (pp. 8.1-8.6).

[10] Zheng, F., Zhang, G., & Song, Z. (2001). Comparisson of different implementation of MFCC. *J. Computer Science & Technology*, 582-589.