Pertanika J. Sci. & Technol. 25 (S): 1 - 10 (2017)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Review Article

Improved Architecture of Speaker Recognition Based on Wavelet Transform and Mel Frequency Cepstral Coefficient (MFCC)

Nor Ashikin Rahman*, Noor Azilah Muda and Norashikin Ahmad

Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia

ABSTRACT

Combining Mel Frequency Cepstral Coefficient with wavelet transform for feature extraction is not new. This paper proposes a new architecture to help in increasing the accuracy of speaker recognition compared with conventional architecture. In conventional speaker model, the voice will undergo noise elimination first before feature extraction. The proposed architecture however, will extract the features and eliminate noise simultaneously. The MFCC is used to extract the voice features while wavelet de-noising technique is used to eliminate the noise contained in the speech signals. Thus, the new architecture achieves two outcomes in one single process: ex-tracting voice feature and elimination of noise.

Keywords: Mel frequency cepstral coefficient, Speaker recognition, Wavelet transform

INTRODUCTION

Human voice consists of unique anatomical structure that can be used to identify a person. Human voice data also contains other elements such as environmental sounds, music, or surroundings discussions which are termed as noises in signals. This has to be eliminated in order to acquire quality and significant voices sounds to recognise a person. Combining two different techniques for these purposes is not new. Yadav and Bhalke, (2015); Maged, Abou El-Farag, and Mesbah, (2014); Sabi-tha and Janardhanan (2013); Abdalla and Ali (2010); and Shafik, Elhalafawy, Diab, Sallam, and El-Samie, (2009) have shown that the combination of MFCC with wave-let transform can improve accuracy. In this paper, a new architecture of speaker recognition is proposed, and which is based on MFCC and wavelet transform.

The paper is organised as follows: Section 2 describes the structure of speaker

ARTICLE INFO

Article history:
Received: 15 August 2016
Accepted: 18 May 2017

E-mail addresses:
luvly.sheqeen@gmail.com (Nor Ashikin Rahman),
azilah@utem.edu.my (Noor Azilah Muda),
norashikin@utem.edu.my (Norashikin Ahmad)
*Corresponding Author

ISSN: 0128-7680

recognition system and previous related studies. Section 3 discusses in depth the proposed architecture. Last but not least, section 4 summaries and concludes the paper.

Speaker Identification System

When individuals speak, they create vibration in the air that can be heard as sound. These vibrations are known as sound waves, and like fingerprints, everyone have their own unique sound waves. However, computers do not understand sound waves and hence, the conversion of sound waves into digital data is vital. This conversion can be done by using an analogue-to-digital converter (ADC). This digital data is then analysed and used further in voice recognition system. Speaker recognition sys-tems consist of two phases: Identification and Verification (Singh, Khan, & Shree, 2012).

During the identification phase, features are extracted from the input speech signal to represent the unique characteristic of individual voice that will be used in creating a speaker model. The obtained voice features will be compared with other features that are stored in a voice model database to identify similarities. If there is no match, the obtained features are stored in the database as a new speaker id. During the veri-fication phase, the features are compared with existing voice features stored in the database and the similarity score is calculated. Approval or rejection of the speaker depends on the similarity score. The speaker is approved if the similarity score is above the threshold. Figure 1 and 2 describe both phases.

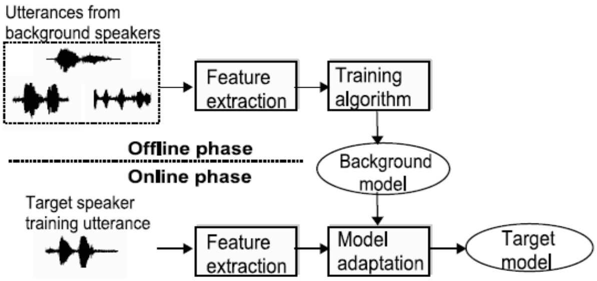

Figure 1. Enrolment or training phase (Adapted from Kinnunen & Li, 2010)

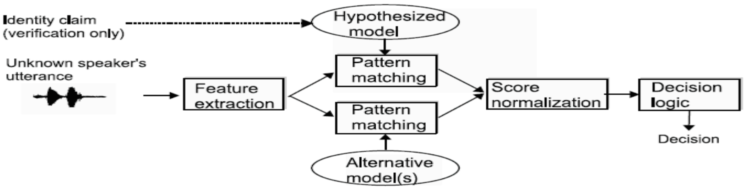

Figure 2. Verification Phase (Adapted from Kinnunen & Li, 2010)

Previous Studies

Maged, Abou El-Farag and Mesbah, (2014) had proposed a robust speaker identifica-tion method from degraded noisy speech signals based on MFCC and Discrete Wavelet Transform. The process of speaker recognition began with noise elimination and after undergoing a few related process, the speech features are then extracted. Both wavelet and MFCC were extracted and used in the study. A wavelet from Daubechies family was used. Noisy signal using Additive White Gaussian Noise (AWGN) was evaluated with different values of Signal-to-Noise Ratio (SNR) (0db, 10db, 20db, 25db, 28db and 30 db) for 8 and 13 speakers. The result showed that the recognition rate improved when compared with MFCC. However, the recognition rate decreased as the number of speakers increased.

Yadav and Bhalke (2015) also studied speaker recognition based on MFCC and wavelet transform using Daubechies 5. Their experiment began with noise elimina-tion, followed by extracting both wavelet and MFCC features. The results of the study showed that the recognition rate obtained using the wavelet-based MFCC was higher than that of traditional MFCC. The results were similar to Maged, Abou El-Farag and Mesbah, (2014) where the recognition rate decreased as the number of speakers increased.

Sabitha and Janardhanan, (2013) also proposed a wavelet based MFCC approach. In this work, a novel family of windowing was used to compute MFCC. Similar to previous studies, both wavelets type Daubechies 4 and MFCC features were extract-ed to gain more features.

Abdalla and Ali (2010) focused on wavelet-based MFCC to form a robust feature extraction algorithm for speech signal. The wavelet transform was used to obtain approximation and detailed resolution channels. The MFCC was later extracted. The identification rate using this method is 97.3%, slightly higher than traditional MFCC which is 93.3% below the 20db noisy signal (Abdalla & Ali, 2010).

Shafik et al., (2009) examined a robust speaker identification method from degrad-ed speech. The proposed wavelet-based MFCC and the traditional feature extraction of MFCCs from noisy speech signals and telephone degraded speech signals with Additive White Gaussian Noise (AWGN) and coloured noise were compared. The results showed that the method proposed by Shafik et al., (2009) improved the recognition rates at different degradation cases. Table 1 shows the summary of pre-vious related studies.

The disadvantages of previous design and implementation work of speaker recog-nition work are: the traditional MFCC for feature extraction is sensitive to noisy en-vironment and the speech signal is said to be stationary. In order to remove the un-wanted noise in speech signal, wavelet transform should be used. Wavelet transform also helps in good representation of stationary as well as non-stationary segments of the speech signal. A review of previous studies showed that all studies extracted both MFCC and wavelet features and combined them to form a large feature vectors. For this work, wavelet features will not be extracted, instead only MFCC features are extracted. Wavelet de-noising technique will be applied to eliminate the noise contained in the speech signal to produce a clean speech features for speaker recognition. The recognition rate obtained by the proposed method will be compared with the previous ones. Table 1 shows the summary of previous studies.

Table 1
Summary of previous studies

Author	Year	Feature	Objective	Recognition rate	Feature Matching	Comment
Yadav and Bhalke, (2015)	2015	MFCC + WAVELET	Speaker identification system based on wavelet transform.	SPEAKER NO. 5 10 15 PROPOSED 92% 88% 77% MFCC 90% 84% 74%	VQ	-Text dependent. Can be improved by testing on Text independent speaker recognition.
Maged, Abou El-Farag, and Mesbah, (2014)	2014		A robust speaker identification method from degraded speech signal.	SPEAKER NO. 8 13 PROPOSED 73.2% 67% MFCC 35.7% 45% (Based on average)	VQLBG	
Sabitha and Janardhanan, (2013);	2013		A novel family of windowing is used to compute MFCC.	MFCC MFCC+DWT CLEAN 98% 100% 10% NOISE 82% 96% 20% NOISE 8% 90% (Based on average)	GMM	-Small amount of speaker dataset. This can be improved by increasing the number of speaker.
Abdalla and Ali, (2010).	2010		Enhanced the performance of MFCC based method in the presence of noise.			
Shafik et al., (2009)	2009		Robust feature extraction algorithm for speech signal.	CLEAN proposed: 99.3% MFCCC: 98.7% NOISY proposed: 97.3% MFCC: 93.3%	HMM	-The recognition rate become lower as the amount of speaker increase. Can be improved by modifying the method so that the recognition rate obtained is higher even in large speaker database. .
Our proposed method	2016		A robust speaker identification method from degraded speech signal.		ANN	
			Text Independent Speaker recognition system based on wavelet transform.		HMM	-Text independent -Only MFCC features are extracted -wavelet of hard and soft thresholding is applied to remove noise. -Applied in larger speaker dataset

Proposed Architecture

The conventional architecture of speaker recognition is based on eliminating the noise first, before proceeding with feature extraction process. This study as mentioned earlier proposes a new architecture (see Figure 3) where the noise is eliminated and features are extracted simultaneously. Its recognition rate is compared with the conventional architecture for contribution to future research. This study is based on Wavelet transform and MFCC whereby the latter's features are extracted from the input speech while wavelet transform is used to suppress noise available in the input speech signal. The steps in extracting the MFCC are described in Figure 4.

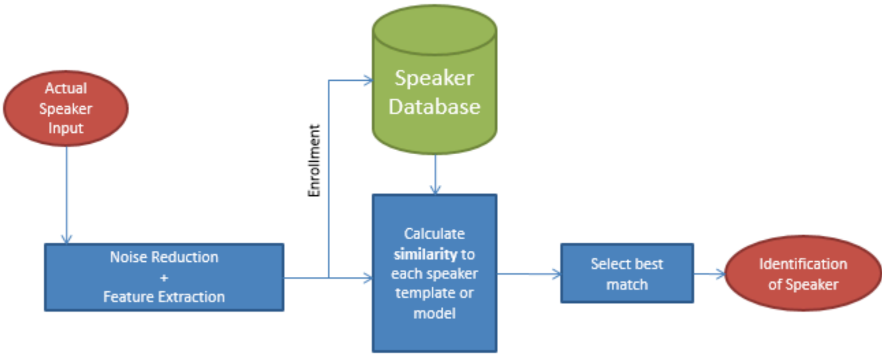
Actual Speaker Input
Speaker Database
Enrollment
Noise Reduction + Feature Extraction
Calculate similarity to each speaker template or model
Select best match
Identification of Speaker

Figure 3. The proposed architecture

Feature Extraction

In this paper, MFCC features are extracted from speech signals through cepstral analysis. The voice input normally recorded at sampling rate is more than 10000Hz. This sampling frequency is chosen to reduce the effect of aliasing during analog-to signal conversion. Figure 4 shows the steps in extracting the MFCC features.

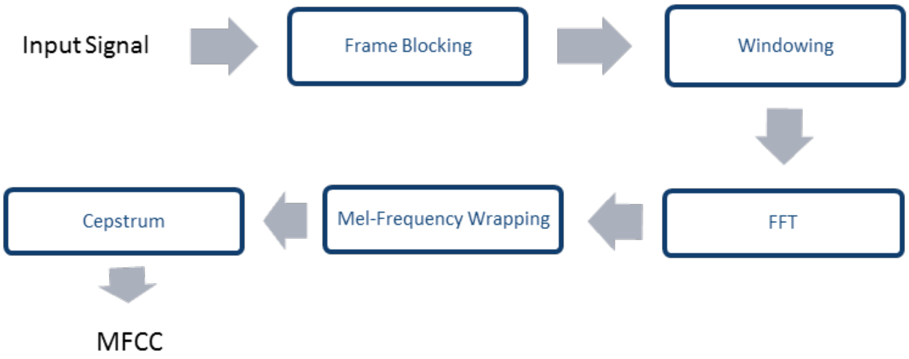

Figure 4. MFCC processor

Step 1: Frame Blocking

From previous research, it was found that speech signals remain stationary for a short period of time (10-30msec) but changes after a long period of time. Thus, the speech signal need to be converted into a number of small frames having frame size N and separated from adjacent frame M (M<N) for further processing (Muda, Begam, & Elamvazuthi, 2010). Usually, the value for N is 256 and M is 100 where N>M (Kin-nunen & Li, 2010). The rate of overlapping between frames is between 35% and 75% (Bharti & Bansal, 2015). The region of overlapping between N and M is calcu-lated by (N-M). Figure 5 shows the overlapping of frames N and M.

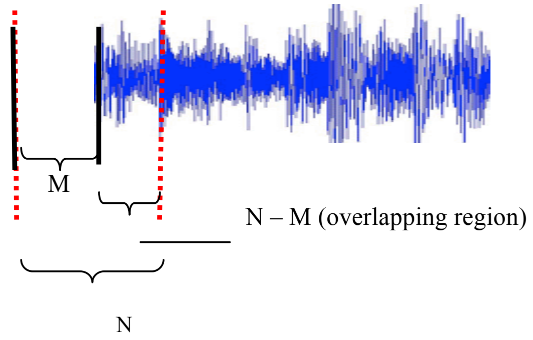

Figure 5. Frame blocking

Step 2: Windowing

Most of the types of windows used are Hamming and Triangular. Hamming window is used to discard the effect of discontinuities at edges of the frame. The equation of hamming window is as follows (Bharti & Bansal, 2015):

$$Y[n] = X[n]*W[n] \tag{1}$$

where Y[n] is output signal, X[n] is input signal and W[n] is hamming window.

$$W[n]=0.54-0.46\cos[2* *n/N-1] \text{ where } 0 \ n \leq N-1 \tag{2}$$

where N is the number of sample in each frame.

Step 3: Fast Fourier Transform (FFT)

This step converts N samples from time domain into the frequency domain. This step is used to eliminate the redundant mathematical calculation and analyse the spectral properties of a signal.

Step 4: Mel Frequency Wrapping

First of all, the N samples need to be converted from time domain to the frequency domain. This step is used to eliminate the redundant mathematical calculations and enable analysing the spectral properties of the signal.

Mel is a unit of measure of perceived pitch/frequency of the tone. It is used be-cause the human perception of the frequency content of acoustic signal does not follow linear scale, instead, it follows the Mel scale. The Mel-frequency is said to be linear frequency spacing when it is below 1000Hz and if it is above 1000Hz, the Mel frequency is logarithmic spacing and it has less details of speech characteristic (as more details are given to lower frequency). The pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold is defined as 1000 Mel. Therefore, the following approximate formula can be used to compute the Mel for a given frequency, f in Hz:

$$\text{Mel}(f) = 2595 * \log10(1 + f/700) \tag{3}$$

In order to stimulate the subjective spectrum, a filter bank is used, one filter for each desired Mel-frequency component (Figure 6). Each filter bank consists of trian-gular bandpass frequency response which is applied in the frequency domain for an efficient result. Overlapping of histogram bins are usually implemented to provide representation of Mel Wrapping in the frequency domain.

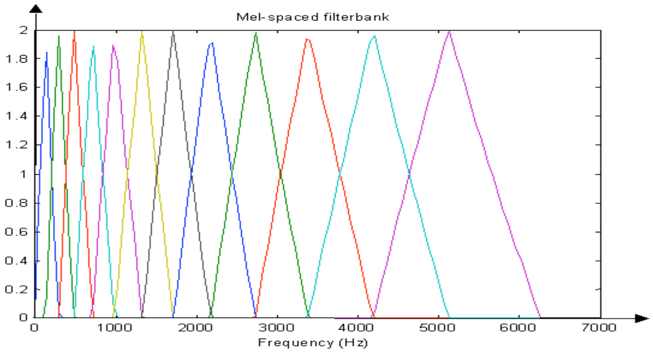

Figure 6. Example of Mel-spaced filter bank

Step 5: Cepstrum

The real number of mel spectrum is converted back to the time domain using Dis-crete Cosine Transform (DCT) to provide better representation of the spectral proper-ties for a given time analysis. The final output of this stage is Mel Frequency Cepstral Coefficient (MFCC).

Noise Elimination

Since MFCC is sensitive to noisy environment, wavelet transform is believed to as-sist in overcoming this weakness. Wavelet de-noising is an operation where noise is eliminated from the noisy speech signals. A threshold value that is a large multiple of the standard deviation of the noise in the speech signal is chosen. Most of the noises are eliminated by thresholding the detail coefficients of the wavelet transformed speech signal. There are two types of thresholding that widely used: hard thresholding and soft thresholding (Govindan, Duraisamy, & Yuan, 2014). The equation of the hard and soft threshold is as below (El-Samie, 2011):

$$f_{hard}(x_w) = \begin{cases} x_w, & |x_w| \geq TH \\ 0, & |x_w| < TH \end{cases} \qquad (1)$$

$$f_{soft}(x_w) = \begin{cases} x_w, & |x_w| \geq TH \\ 2x_w - TH, & \frac{TH}{2} \leq x_w < TH \\ TH + 2x_w, & -TH < x_w \leq -\frac{TH}{2} \\ 0, & |x_w| < \frac{TH}{2} \end{cases} \qquad (1)$$

where TH denotes the threshold value while x_w denotes the coefficient of the high frequency components of the DWT.

Feature Matching

The objective of feature matching is to differentiate between one speaker from another. For this purpose, the match score is calculated by measuring the similarity between the feature vectors of the input voice and models, template model and stochastic models. In template models, the pattern matching is deterministic. In order to minimise a distance measure value, the alignment of the observed frames to template frames is selected. While for stochastic models, the pattern matching is probabilistic. It is a measure of the likelihood, or conditional probability, of the observation given the model (Jain, Bolle, & Pankanti, 2006). Dynamic Time Warping (DTW) is one of the template models while Hidden Markov Models (HMMs) is one of the stochastic models. Normally, HMM is the most used model compared with template models due to its flexibility which allow using speech units from sub-phoneme units to words and enabling the design of text-prompted systems (Campbell, 1995). The present study uses HMM for feature matching.

CONCLUSION

There is still a room for improving the performance of MFCC in noisy speech signal. This paper proposed wavelet based MFCC method for speaker recognition where the noise is eliminated while extracting the speech features at the same time. The previ-ous studies eliminated the noise first followed by feature extraction. In this work, wavelet was used to eliminate noise in speech signal by applying soft and hard threshold. The MFCC features were extracted from the clean speech signal. The recognition rate obtained was compared with the previous studies which are used conventional architecture.

ACKNOWLEDGEMENT

This work was funded by the Ministry of Higher Education Malaysia and Universiti Teknikal Malaysia Melaka (UTeM) through the Fundamental Research Grant Scheme - FRGS/2/2013/ICT02/FTMK/02/5/F00188

REFERENCES

Abdalla, M. I., & Ali, H. S. (2010). Wavelet-based mel-frequency cepstral coeffi-cients for speaker identification using hidden markov models. *arXiv preprint arXiv:1003.5627*.

Bharti, R., & Bansal, P. (2015). Real time speaker recognition system using MFCC and vector quantization technique. *International Journal of Computer Applications, 117*(1).

Campbell Jr, J. P. (1995, May). Testing with the YOHO CD-ROM voice verification corpus. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference*, (Vol. 1, pp. 341-344). IEEE.

El-Samie, F. E. A. (2011). *Information security for automatic speaker identification* (pp. 1-122). Springer New York.

Govindan, S. M., Duraisamy, P., & Yuan, X. (2014). Adaptive wavelet shrinkage for noise robust speaker recognition. *Digital Signal Processing, 33*, 180-190.

Jain, A., Bolle, R., & Pankanti, S. (Eds.). (2006). *Biometrics: personal identification in networked society* (Vol. 479). Springer Science & Business Media.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication, 52*(1), 12-40.

Maged, H., Abou El-Farag, A., & Mesbah, S. (2014, June). Improving speaker identi-fication system using discrete wavelet transform and AWGN. *IEEE International Conference in Software Engineering and Service Science (ICSESS) (*pp. 1171-1176). IEEE.

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083*.

Sabitha, V., & Janardhanan, P. (2013, January). Performance analysis of speaker edentification system using MFCC and DWT under various noise levels. In *International Journal of Engineering Research and Technology, 2*(6), (June-2013). ESRSA Publications.

Shafik, A., Elhalafawy, S. M., Diab, S. M., Sallam, B. M., & El-Samie, F. A. (2009). A wavelet based approach for speaker identification from degraded speech. *International Journal of Communication Networks and Information Security (IJCNIS), 1*(3).

Singh, N., Khan, R. A., & Shree, R. (2012). Applications of Speaker Recogni-tion. *Procedia Engineering, 38*, 3122-3126.

Yadav, S. S., & Bhalke, D. G. (2015). Speaker identification system using wavelet transform and VQ modeling technique. *International Journal of Computer Applications, 112*(9).