Significance of Frequency Band Selection of MFCC for Text-Independent Speaker Identification

S.B. Dhonde and S.M. Jagade

Abstract This paper presents significance of Mel-frequency Cepstral Coefficients (MFCC) Frequency band selection for text-independent speaker identification. Recent studies have been focused on speaker specific information that may extends beyond telephonic passband. The selection of the frequency band is an important factor to effectively capture the speaker specific information present in the speech signal for speaker recognition. This paper focuses on development of a speaker identification system based on MFCC features which are modeled using vector quantization. Here, the frequency band is varied up to 7.75 kHz. Speaker identification experiments evaluated on TIMIT database consisting of 630 speaker shows that the average recognition rate achieved is 97.37 % in frequency band 0–4.85 kHz for 20 MFCC filters.

Keywords Speaker recognition · Feature extraction · Mel scale · Vector quantization

1 Introduction

Speaker recognition is nothing but to recognize the person from known set of voices. Speaker recognition is classified into speaker identification and speaker verification. Speaker identification is nothing but to identify a person from the known set of voices. It is a task of identifying who is talking from known set of voice samples. While, speaker verification is to verify claimed identity of a speaker,

S.B. Dhonde (✉)
Department of Electronics Engineering, All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune 411001, India
e-mail: dhondesomnath@gmail.com

S.M. Jagade
Department of Electronics and Telecommunication Engineering, TPCT COE, Osmanabad 413501, India
e-mail: smjagade@yahoo.co.in

S.C. Satapathy et al. (eds.), *Proceedings of the International Conference on Data Engineering and Communication Technology*, Advances in Intelligent Systems and Computing 469, DOI 10.1007/978-981-10-1678-3_21

i.e., Yes or No decision. Speaker identification is further classified into text-dependent identification and text-independent identification. Text-dependent speaker identification requires same utterance in training and testing phase. Whereas, in text-independent speaker identification training and testing utterances are different. Speaker identification system consists of two distinct phases, a training phase and testing phase. In training phase, the features computed from voice of speaker are modeled and stored in the database. In testing phase, the features extracted from utterance of unknown speaker are compared with the speaker models stored in database to identify the unknown person.

Feature extraction step in speaker identification transforms the raw speech signal into a set of feature vectors. The raw speech signal is represented in compact, less redundant feature vectors [1]. Features emphasizing on speaker specific properties are used to train speaker model. As feature extraction is the first step in speaker identification, the quality of the speaker modeling and classification depends on it [2].

In the computation of MFCCs, the spectrum is estimated from windowed speech frames. The spectrum is then multiplied by triangular Mel filter bank to perform auditory spectra analysis. Next step is the logarithm of windowed signal followed by discrete cosine transform. An important step in the computation of MFCC is the Mel filter bank [1, 3]. The MFCC technique computes speech parameters based on how human hears and perceives sound [2]. However, MFCC does not consider the contribution of piriform fossa, which results in high frequency components [4].

The auditory filter created by the cochlea inside human ear has frequency bandwidth termed as critical band. The existence of auditory filter is experimented by Harvey Fletcher [5]. The auditory filters are responsible for frequency selectivity inside the cochlea which helps the listener for discrimination between different sounds. These critical band filters are designed using frequency scales, i.e., the Mel scale and the Bark scale [5]. The MFCCs are widely used in speaker recognition system [2, 6–8]. In previous work, many researchers have demonstrated the dominant performance of MFCCs and contributed to enhance the robustness of MFCC features as well as speaker recognition system. Such efforts are [2, 7–15]. The importance of speaker specific information present in the wideband speech in demonstrated in [16].

This paper presents the importance of frequency band selection. The speaker specific information extends beyond telephonic pass band [16]. The performance of MFCC scheme in different frequency bands is demonstrated in this paper. The organization of this paper is as follows. Section 2 discusses frequency warping scale Mel scale and MFCC computation process. Experimental set-up is discussed in the Sect. 3. Section 4 discusses the results followed by conclusion in Sect. 5.

2 Mel Scale and MFCC

Nerves in human ear perception system responds differently to various frequencies in a listened sound. For example, sound of 1 kHz triggers nerves while sound of other frequencies will keep quite. This scale is roughly nonlinear in nature. It is like a band-pass filter that looks like triangular in shape. This was observed for how human ear perceives Melody sound. Mel scale is based on pitch perception [5]. Mel scale uses triangular-shaped filters and is roughly linear below 1 kHz and logarithmically nonlinear above 1 kHz. The relationship between Mel scale frequencies and linear frequencies is given as per the following equation,

$$F_{mel} = 2595^{*} \log_{10}\left(1 + \frac{F_{Linear}}{700}\right) \quad (1)$$

Figure 1 shows Mel scale filter bank. MFCC procedure starts with pre-emphasis which boosts the higher frequencies. The high-pass filter given by transfer function, $H(z) = 1 - az^{-1}$ where, $0.9 \leq a \leq 1$ is generally used for pre-emphasis. The pre-emphasized signal is divided into frames of duration 10–30 ms with 25–50 % overlap to avoid loss of information. Over this short duration, speech signal is assumed to remain stationary. Then, each frame is multiplied with Hamming window in order to smooth the speech signal. After windowing step, fast Fourier transform is used to estimate the frequency content present in speech signal. Next, the windowed spectrum is integrated with Mel filter bank which is based on Mel scale as given in Eq. (1). The vocal tract response is separated from excitation signal using logarithm of windowed spectrum integrated with Mel filter bank followed by discrete cosine transform.

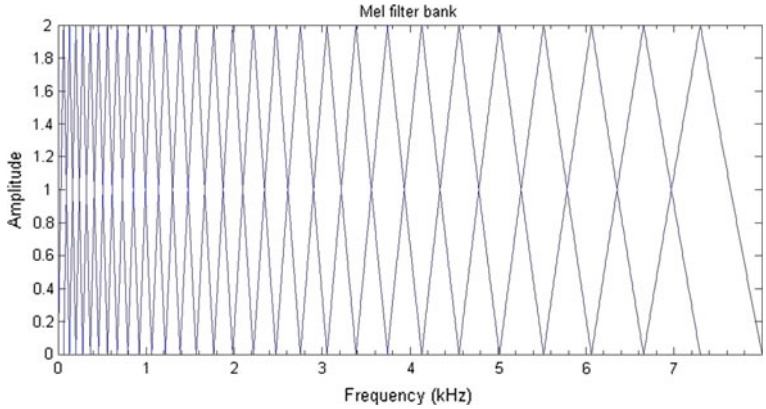

Fig. 1 Mel filter bank

3 Experimental Set-up

In this paper, the performance of Mel-frequency cepstral coefficients frequency band selection for text-independent speaker identification system is evaluated on TIMIT [17] database. TIMIT database consists of a total number of recordings of 630 speakers among which 438 are male speakers and 192 are female speakers. There are ten different sentences of each speaker of sampling frequency 16 kHz which makes a total of 6300 sentences recorded from 8 dialect region of the United States. For training of speaker model, eight sentences, five SX and three SI (approximately 24 s) were used. For testing purpose, two remaining SA sentences (sentences of 3 s each) were used. All the experiments have been performed using HP Pavilion g6 laptop with CPU speed of 2.50 GHz, 4 GB RAM, and MATLAB 8.1 signal processing tool.

The speech signal has been pre-emphasized with the first-order high pass filter given by equation $H(z) = 1–0.95z^{-1}$. The signal is divided into 256 samples per frame with 50 % overlap followed by the Hamming window. The spectrum of the windowed signal is calculated by fast Fourier transform (FFT). The spectrum is multiplied by Mel-filter bank followed by logarithm and discrete cosine transform (DCT) to obtain MFCCs. Speaker model is generated for each speaker from the MFCCs using vector quantization (LBG algorithm). This speaker model is stored in the database. In testing phase, MFCC features of an unknown speaker are extracted. Next, Euclidean distance between MFCC features and speaker model stored in the database is calculated. The speaker is recognized on the basis of minimum Euclidean distance computed between MFCC features in testing phase and speaker model stored in database. The experiments are carried out for different number of MFCC filters, i.e., 20 and 29 in the frequency band 0–4 kHz. Next, frequency is varied up to 7.75 kHz with 20 MFCC filters. The number of MFCC filters is varied as 13, 20, and 29 in the significant frequency band to observe the average recognition rate. In each experiment, first 12 cepstral coefficients excluding the 0th coefficient are selected and the number of clusters of vector quantization is 32.

4 Results and Discussion

Frequency band 0–4 kHz is analyzed for MFCC filters equal to 20 and 29. This frequency band is analyzed in two separate intervals. First, frequency band 0–2 kHz is analyzed and then frequency band 0–4 kHz is analyzed. The recognition rate in percentage is calculated by,

$$\text{Recognition Rate} = \frac{\text{Number of correct matches}}{\text{Total number of test speaker}} \times 100\%$$

Table 1 Recognition rate for frequency range 0–4 kHz

Sr. no.	Frequency band (kHz)	Sampling frequency (kHz)	No. of filters	Average recognition rate
1	0–2	0–4	20	81.18
2	0–2	0–4	29	83.41
3	0–4	0–8	20	95.95
4	0–4	0–8	29	95.80

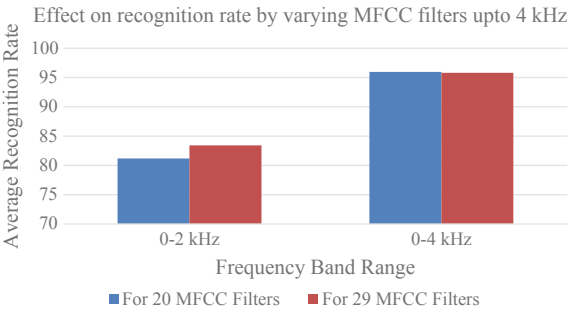

Fig. 2 Effect on recognition rate by varying MFCC filters up to 8 kHz

The following Table 1 and Fig. 2 shows the average recognition rate observed in these bands.

It is observed that the frequency band 0–4 kHz has provided a good resolution as compared to frequency band 0–2 kHz. This is because average recognition rate of 95.95 % is observed for 20 MFCC filters in frequency band 0–4 kHz. This indicates that speaker specific information is present up to 4 kHz. Also, varying the number of filters in these bands has less effect on recognition rate as compared to variation in frequency band. In addition to number of filters, it is also important to select a frequency band which is having good resolution for speaker identification.

In next subsequent experiments 20 MFCC filters are chosen and frequency band is varied up to 7.75 kHz. Table 2 and Fig. 3 shows the effect on recognition rate by varying frequency band.

Table 2 Effect on average recognition rate by varying frequency band up to 7.75 kHz for 20 MFCC filters

Sr. no.	No. of filters	Frequency band (kHz)	Sampling frequency (kHz)	Average recognition rate
1	20	0–4	0–8	95.95
2	20	0–4.75	0–9.5	96.66
3	20	0–4.85	0–9.7	97.37
4	20	0–4.9	0–9.8	96.90
5	20	0–4.95	0–9.9	96.58
6	20	0–5	0–10	96.66
7	20	0–6	0–12	81.58
8	20	0–7.75	0–15.5	61.83

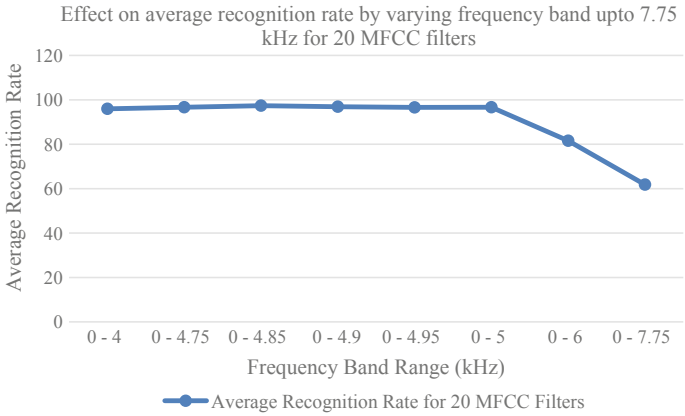

Fig. 3 Effect on average recognition rate by varying frequency band for 20 MFCC filters

Table 3 Effect of varying MFCC filters in frequency band 0–4.85 kHz

Sr. no.	No. of MFCC filters	Frequency band (kHz)	Sampling frequency (kHz)	Average recognition rate
1	13	0–4.85	0–9.7	95.32
2	20	0–4.85	0–9.7	97.37
3	29	0–4.85	0–9.7	97.30

From Table 2, it is observed that frequency band 0–4.85 kHz is the significant frequency band. This is because the maximum average recognition rate achieved is 97.37 % in this frequency band for 20 MFCC filters. Thereafter, average recognition rate is decreasing as shown in Table 2. It is observed that speaker specific information extends beyond 4 kHz, and therefore, it is important to select frequency band. Next, in the significant frequency band, i.e., 0–4.85 kHz, MFCC filters are varied and effect on average recognition rate is observed. Following table shows the effect of varying MFCC filters in frequency band 0–4.85 kHz.

From Table 3 and Fig. 4, it is observed that there is no much more improvement in average recognition rate by varying number of filters in the significant frequency band.

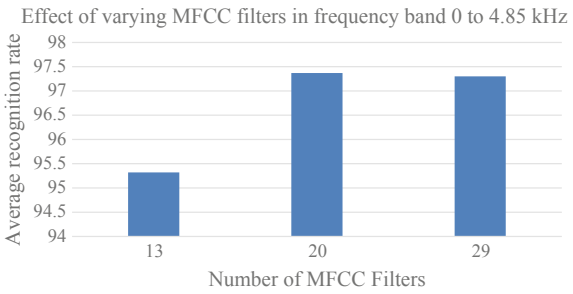

Fig. 4 Effect of varying MFCC filters in frequency band 0–4.85 kHz

5 Conclusion

In this paper, the significance of selection of Mel-frequency Cepstral Coefficients (MFCC) frequency band for speaker identification is proposed. First, frequency band 0–2 kHz is selected and MFCC filters are varied in this frequency band. Next, frequency band is varied 0–4 kHz and MFCC filters are varied in this band. It is found that speaker specific information is present in the frequency band 0–4 kHz is much more as compared to 0–2 kHz. Further, frequency band is varied up to 7.75 kHz. It is observed that the average recognition rate achieved is 97.37 % in the frequency band 0–4.85 kHz for 20 MFCC filters. This indicates that speaker specific information is present up to 4.85 kHz. Thereafter, recognition rate is decreasing. In the significant frequency band 0–4.85 kHz, MFCC filters are varied as 13, 20, and 29 and it is observed that there is no much more improvement in the average recognition rate.

References

1. Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, Douglas A. Reynolds: A tutorial on text-independent speaker verification, EURASIP Journal on Applied Signal Processing 2004, Hindawi, pp. 430–451 (2004).
2. Md Jahangir Alam, Tomi Kinnunen, Patrick Kenny, Pierre Ouellet, Douglas O'Shaughnessy: Multitaper MFCC and PLP features for speaker verification using i-vectors, Journal on Speech Communication, Elsevier, vol. 55, no. 2, pp. 237–251 (2013).
3. Claude Turner, Anthony Joseph, Murat Aksu, Heather Langdond: The Wavelet and Fourier Transforms in Feature Extraction for Text-Dependent, Filterbank-Based Speaker Recognition, Journal onProcedia Computer Science, Elsevier, vol. 6, pp. 124–129 (2011).
4. Mangesh S. Deshpande, Raghunath S. Holambe: New Filter Structure based Admissible Wavelet Packet Transform for Text-Independent Speaker Identification, International Journal of Recent Trends in Engineering, vol. 2, no. 5, pp. 121–125 (2009).
5. Dr. Shaila D. Apte: Speech Processing Applications, in Speech and Audio Processing, Section 1, Section 2 and Section 3, pp. 1–6, 67, 91–92, 105–107, 129–132, Wiley India Edition.
6. Tomi Kinnunen, Haizhou Li: An overview of text-independent speaker recognition: From features to supervectors, Journal onSpeech Communication, Elsevier, vol. 52, no. 1, pp. 12–40 (2010).
7. Tomi Kinnunen, Rahim Saeidi, FilipSedlák, Kong Aik Lee, Johan Sandberg, Maria Hansson-Sandsten, Haizhou Li: Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification, IEEE Transactions Audio, Speech and Language Processing, vol.20, no.7, pp. 1990–2001 (2012).
8. Pawan K. Ajmera, Dattatray V. Jadhav, Ragunath S. Holambe: Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram, Journal on Pattern Recognition, Elsevier, vol. 44, no. 10–11, pp. 2749–2759 (2011).
9. WU Zunjing, CAO Zhigang: Improved MFCC-Based Feature for Robust Speaker Identification, TUP Journals & Magazines, vol.10, no 2, pp. 158–161 (2005).

10. Jian-Da Wu, Bing-Fu Lin: Speaker identification using discrete wavelet packet transform technique with irregular decomposition, Journal on Expert Systems with Applications, Elsevier, vol. 36, no. 2, pp. 3136–3143 (2009).
11. R. Shantha Selva Kumari, S. Selva Nidhyananthan, Anand.G: Fused Mel Feature sets based Text-Independent Speaker Identification using Gaussian Mixture Model, International Conference on Communication Technology and System Design 2011, Journal on Procedia Engineering, Elsevier, vol. 30, pp. 319–326 (2012).
12. Seiichi Nakagawa, Longbiao Wang, and Shinji Ohtsuka: Speaker Identification and Verification by Combining MFCC and Phase Information, IEEE Transactions Audio, Speech and Language Processing, vol.20, no.4, pp. 1085–1095 (2012).
13. Sumithra Manimegalai Govindan, Prakash Duraisamy, Xiaohui Yuan: Adaptive wavelet shrinkage for noise robust speaker recognition, Journal on Digital Signal Processing, Elsevier, vol. 33, pp. 180–190 (2014).
14. Noor Almaadeed, Amar Aggoun, Abbes Amira: Speaker identification using multimodal neural networks and wavelet analysis, IET Journals and Magazines, vol. 4, no. 1, pp. 18–28 (2015).
15. Khaled Daqrouq, Tarek A. Tutunji: Speaker identification using vowels features through a combinedmethod of formants, wavelets, and neural network classifiers, Journal on Applied Soft Computing, Elsevier, vol. 27, pp. 231–239 (2015).
16. Pradhan, G.; Prasanna, S.: Significance of speaker information in wideband speech, in Communications (NCC), 2011 National Conference on, pp. 1–5, (2011).
17. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, V. Zue, TIMIT acoustic-phonetic continuous speech corpus, http://catalog.ldc.upenn.edu/ldc93s1, 1993.