MFCC Analysis of Biometric Person Identification System in Multilingual Speech Environment

Tushar Kant Sahu
Department of Electronics & Telecommunication Engineering
Shri Shankracharya College of Engineering and Technology bhilai

Vinay Jain
Department of Electronics & Telecommunication Engineering
Shri Shankracharya College of Engineering and Technology bhilai

Abstract

In this paper we explain the multilingual speaker identification system. Speaker identification is conducted on 3 Indian languages (Hindi, Marathi and Rajasthani). We create a database of 25 person in each language. In our system we use 3 different sentences and each sentence in 3 language. We focus on the effect of language mismatch in the speaker identification performance of individual languages and all languages together. Mel Frequency Cepstrum Coefficient (MFCC) is used for feature extraction. The standard SVM-based speaker identification is used.

Keywords- Multilingual Speaker Identification, MFCC, Speaker Identification, Support Vector machine, Indian languages

I. INTRODUCTION

Speaker recognition (SR) is the task of recognizing a person by processing his/her spoken utterances. An ideal SR system is expected to perform effectively irrespective of changes in session, emotion, health and language of a speaker. Prior work demonstrates the prominent effect of spoken languages in text independent SR accuracy. Multilingual speaker recognition has thus been a field of active research in recent years. Apart from the common issues concerning conventional SR systems, a major challenge in this field is the collection of adequate data for preparing a speech corpus. In the context of Indian languages, a suitable corpus should span most of the local languages spoken in the country. The distribution of speakers in each language should be made according to a recent census. Care must be taken to retain accent variations in each language.

II. METHODOLOGY

In this section we discuss the development of the multilingual speaker recognition systems in details.

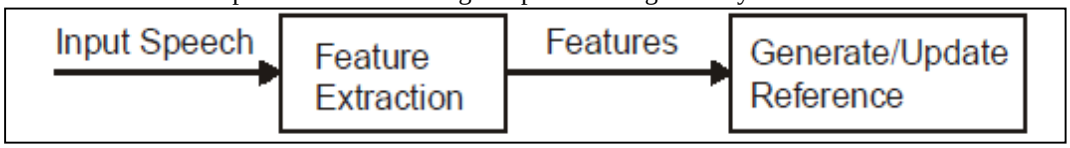

Fig. 1: Block diagram of training phase

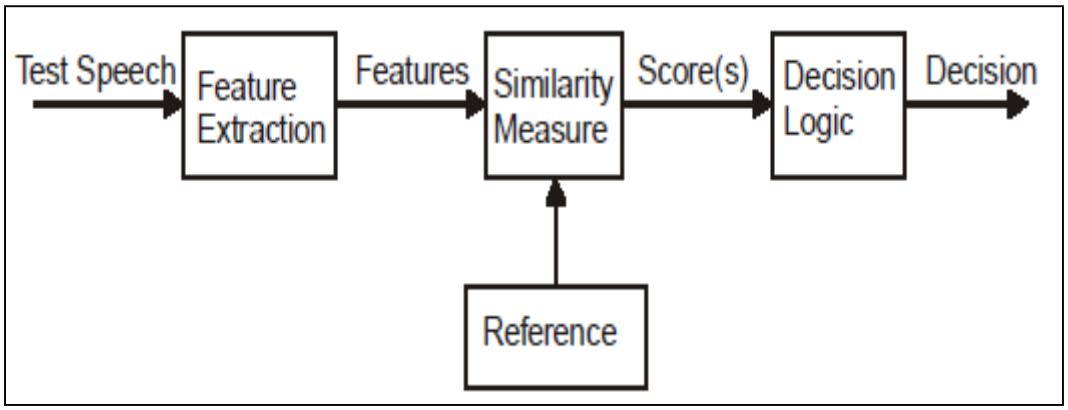

Fig. 2: Block diagram of testing phase

A. Database Generation

We took the voice samples of 25 person for our database. Each person's voice recorded in 3 languages i.e. HIndi, Marathi and Rajasthani at the sampling rate of 48 kHz. Voices are recorded with the help of mobile phone.

B. Feature Extraction

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

The standard implementation of computing the Mel-Frequency Cepstral Coefficients is shown in Figure and the exact steps are described below-

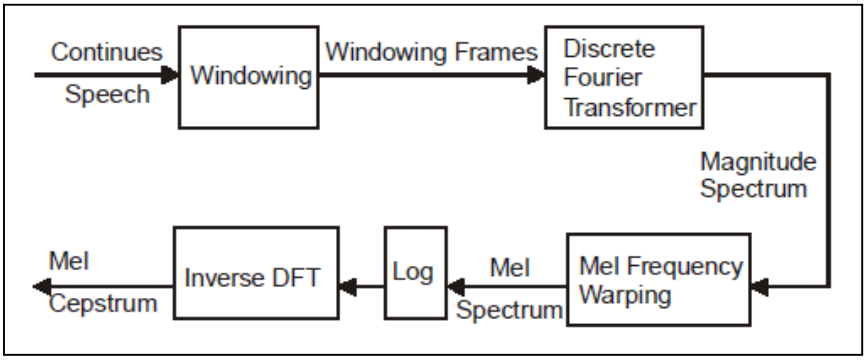

Fig. 3: Block diagram for computing MFCC

1) Pre-Emphasis

It is a filtering technique that stress on the higher frequencies. Some sounds have a steep roll-off in high frequency. So, to balance the speech spectrum of voiced sounds, high-frequency filtering is needed. Pre-emphasis is done by following equation.

$$H(z) = 1 - \alpha z^{-1} \quad (1)$$

Where the value of α controls the slope of the filter and is usually between 0.9 to 1.0.

2) Windowing

Speech is a non-stationary time variant signal. A signal is considered to be stationary if its frequency does not change over time. We assume that human speech is built from a dictionary of phonemes, while for most of the phonemes the properties of speech remain invariant for a short period of time. Thus we assume the signal behaves stationary for those time frames. In order to obtain frames we multiply the speech signal with a windowing function. Hamming window is used for this operation.

3) Discrete Fourier Transform (DFT)

Each windowed frame is converted into magnitude spectrum by applying DFT using the equation 2.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi k}{N}}, \quad 0 \le k \le N-1 \quad (2)$$

Where x(n) is the samples of the windowed signal. X(k) is the magnitude spectrum of Windowed signal and N is the number of points used to compute the DFT.

4) Mel-Spectrum

The Mel-spectrum is computed by passing the DFT spectrum through a set of band-pass triangular filters known as Mel-filter bank. The Mel scale is a mapping between the physical frequency scale (Hz) and the perceived frequency scale (Mels). This can be can be expressed by the following equation.

$$fmel = 2595 \log(1 + f / 700) \quad (3)$$

Where f denotes the physical frequency and fmel denotes the perceived mel-frequency. The mel-spectrum values or mel frequency coefficients of the magnitude spectrum X(k) is computed by multiplying the magnitude spectrum by each of the triangular Mel-weighting filters.

$$S(m) = \sum_{k=0}^{N-1} |X(k)|^2 H_m(k), \quad 0 \le m \le M-1 \quad (4)$$

Where S(m) is mel-frequency coefficients and *M* is total number of triangular mel-weighting filters.

5) Inverse Discrete Cosine Transform (IDCT)

Log operation is performed on the Mel frequency coefficients. The IDCT is used to calculate the cepstral coefficients. MFCC is computed as:

$$c(n) = \sum_{m=0}^{M-1} \log(S(m)) \cos\left(\frac{\pi n(m-0.5)}{M}\right), \quad n = 0,1,2, \ldots\ldots, C-1 \qquad (5)$$

Where c(n) are the cepstral coefficients and *C* is the number of MFCCs.

6) Deltas and Delta-Deltas

Also known as differential and acceleration coefficients. The MFCC feature vector describes only the power spectral envelope of a single frame, but it seems like speech would also have information in the dynamics i.e. what are the trajectories of the MFCC coefficients over time. It turns out that calculating the MFCC trajectories and appending them to the original feature vector increases ASR performance by quite a bit (if we have 12 MFCC coefficients, we would also get 12 delta coefficients, which would combine to give a feature vector of length 24).

To calculate the delta coefficients, the following formula is used:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \qquad (6)$$

Where d_t is a delta coefficient, from frame t computed in terms of the static coefficients C_t+_n to C_{t-n}. A typical value for N is 2. Delta-Delta (Acceleration) coefficients are calculated in the same way, but they are calculated from the deltas, not the static coefficients.

III. RESULT

A database of 25 speakers is created .The feature extraction was done by using MFCC (Mel Frequency Cepstral Coefficients). All of the input speech signals are sampled at 48 kHz. A speaker identification system comprises of a training phase and a test phase. In the training phase the SVM models are created for each speaker. In testing phase the stored data are compared with the claimed SVM model and a decision is made.

In this section the audio samples (in Hindi, Marathi and Rajasthani language) of one speaker is passed through the MFCC pipeline as shown in figure (3) and MFCC values computed. Various plots of these 3 samples are given in below figure. MFCC values obtained from voice sample of one speaker in 3 different languages are as follows-

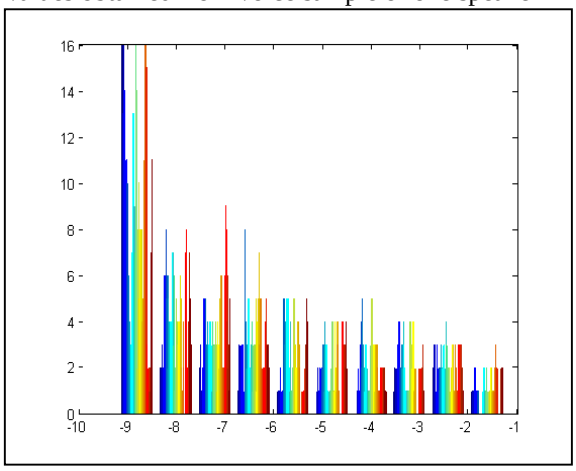

Fig. 4(a):

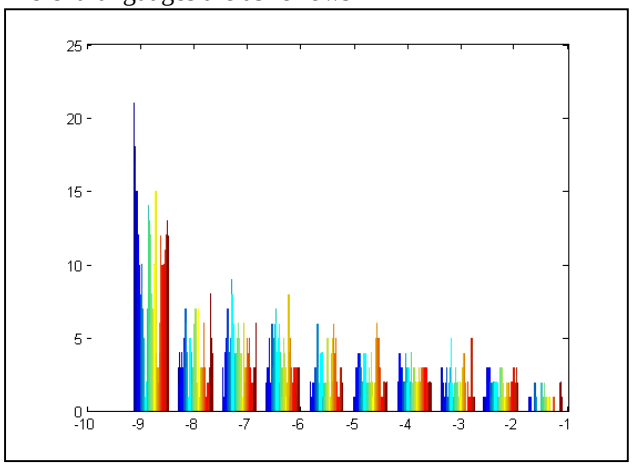

Fig. 4(b):

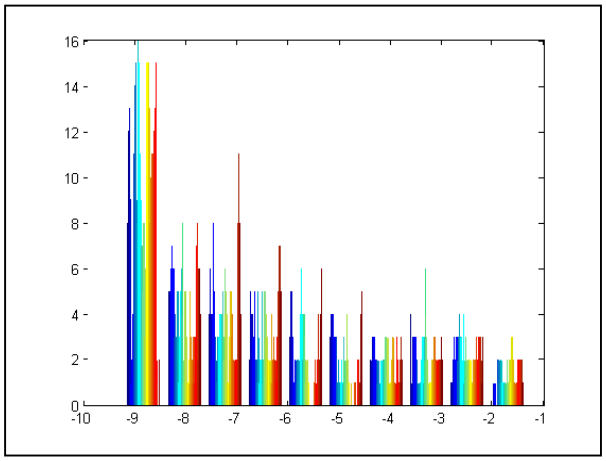

Fig. 4(c):

Fig. 4: MFCC values of a voice sample of 3 languages (a) Hindi, (b) Marathi, (c) Rajasthani

We can see from the above histogram plots, all the MFCC values varies from -10 to -1. Although the audio samples of one speaker is taken but we can see the deviation in values from above plots with respect to language spoken.

IV. CONCLUSION

In this paper we presented a brief idea about multilingual speaker identification system. MFCC values of 3 voice samples of same speaker recorded in 3 different languages. The histogram bar plot of these samples indicate that there are variations in extracted features of speaker's samples of different languages. In future work, we will train and test these MFCC values with the help of Support Vector Machine (SVM) and perform the speaker identification operation.

REFERENCES

[1] Sourjya Sarkar, K. Sreenivasa Rao, Dipanjan Nandi and Sunil Kumar,"Multilingual Speaker Recognition on Indian Languages", Annual IEEE India Conference (INDICON) 2013.
[2] Javier Gonzalez-Dominguez, Member, IEEE, David Eustis, Ignacio Lopez-Moreno, Member, IEEE, Andrew Senior, Senior Member, IEEE, Françoise Beaufays, Senior Member, IEEE, and Pedro J. Moreno, Senior Member, IEEE ,"A Real-Time End-to-End Multilingual Speech Recognition Architecture", IEEE Journal of Selected Topics In Signal Processing, VOL. 9, NO. 4, JUNE 2015.
[3] Michal ADAMSKI, Prof. Basie VON SOLMS,"An Open Speaker Recognition Enabled Identification and Authentication System",IST-Africa 2014 Conference Proceedings Paul Cunningham and Miriam Cunningham (Eds) IIMC International Information Management Corporation, 2014.
[4] Shigeki Matsuda, Xinhui Hu, Yoshinori Shiga, Hideki Kashioka, Chiori Hori, Keiji Yasuda, Hideo Okuma, Masao Uchiyama, Eiichiro Sumita, Hisashi Kawai, and Satoshi Nakamura,"Multilingual speech-to-speech translation system:VoiceTra",IEEE 14th International Conference on Mobile Data Management 2013.
[5] Hui Lin∗†, Jui-ting Huang∗‡, Franc¸oise Beaufays∗, Brian Strope∗, Yun-hsuan Sung," Recognition of Multilingual Speech in Mobile Applications", 978-1-4673-0046-9/12/$26.00 ©2012 IEEE.
[6] U. Bhattacharjee and A. Sarmah, "A multilingual speech database for speaker recognition," in Proc. IEEE International Conference on Signal Processing, Computing and Control (ISPCC), 2012, pp. 1–5.
[7] K. S. Rao, S. Maity, and V. R. Reddy, "Pitch synchronous and glottal closure based speech analysis for language recognition," International Journal of Speech Technology, vol. Springer (Accepted, DOI: 10.1007/s10772-013-9193-5), 2013.
[8] S. Maity, A. Vuppala, K. S. Rao, and D. Nandi, "IITKGP-MLILSC Speech Database for Language Identification," in Proc. IEEE 18th National Conference on Communications, 2012, pp. 1–5.
[9] V. R. Reddy, S. Maity, and K. S. Rao, "Identification of Indian languages using multi-level spectral and prosodic features," International Journal of Speech Technology (Springer), vol. DOI: 10.1007/s10772-013-9198- 0, 2013
[10] K. S. Rao, S. Maity, and V. R. Reddy, "Pitch synchronous and glottal closure based speech analysis for language recognition," International Journal of Speech Technology (Springer), vol. DOI: 10.1007/s10772- 013-9193-5, 2013.
[11] B. Nagaraja and H. Jayanna, "Multilingual Speaker Identification with the Constraint of Limited Data Using Multitaper MFCC," pp. 127–134, 2012.
[12] H. Caesar, "Integrating language identification to improve multilingual speech recognition," Idiap, Idiap-RR Idiap-RR-24–2012, no. 7, 2012.
[13] H. Lin, J. T. Huang, F. Beaufays, B. Strope, and Y. H. Sung, "Recognition of multilingual speech in mobile applications," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Mar. 2012, pp. 4881–4884.
[14] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng, "Large scale distributed deep networks," in Advances in Neural Information Processing Systems 25, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Cambridge, MA, USA: MIT Press, 2012, pp. 1232–1240.
[15] G. Heigold, V. Vanhoucke, A. W. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2013, pp. 8619–8623.
[16] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2014, pp. 5337–5341.
[17] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 14–22, Jan. 2012.
[18] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.
[19] D. Ciresan, U. Meier, L. Gambardella, and J. Schmidhuber, "Deep big simple neural nets excel on handwritten digit recognition," CoRR, vol. Abs/1003.0358, 2010.
[20] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing, Exploratory DSP," IEEE Signal Processing Mag., vol. 28, no. 1, pp. 145–154, Jan. 2011.