



Implementation of Acoustic Phonetic Features for Speaker Verification– A Review

Bidyut Kumar Das*, Utpal Bhattacharjee#

*Department of Computer Sc, M.C.College, Assam, India

#Department of Computer Sc & Engineering, RGU, Arunachal Pradesh, India

I. Introduction

Speaker verification is the computing task of validating a user's claimed identity using characteristics extracted from their voices. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, voice mail etc.

Speaker recognition methods can be divided into text-dependent and text-independent. The former require the speaker to say key words or sentences having the same text for both training and recognition trials, whereas the latter do not rely on a specific text being spoken.

Both text-dependent and text-independent methods however share a common problem. These systems can be easily deceived because someone who plays back the recorded voice of a registered speaker saying the keywords or sentences can be accepted as the registered speaker. To cope with this problem, there are methods in which a small set of words, such as digits, are used as keywords and each user is prompted to utter a given sequence of key words that is randomly chosen every time the system is used.

There are two major applications of speaker recognition technologies and methodologies. If the speaker claims to be of a certain identity and the voice is used to verify this claim this is called verification or authentication. On the other hand, identification is the task of determining an unknown speaker's identity. In other words speaker verification is a 1:1 match where one speaker's voice is matched to one template (also called a "voice print" or "voice model") whereas speaker identification is a 1: N match where the voice is compared against N templates.

II. Features

Mel Frequency Cepstral Coefficients (MFCC):

The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. The formula for converting from frequency to Mel scale is:

$$M(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (1)$$

A **cepstrum** is the result of taking the Fourier transform (FT) of the logarithm of the estimated spectrum of a signal. 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").

MFCCs are commonly derived as follows:

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

Author of [1] use MFCC with Vector Quantization for speaker recognition. Emphasize is given on number of filter and type of window (Hanning and Rectangular). In signal processing, window function is a mathematical function that is zero-valued outside of some selected interval. For example, a function that is constant inside the interval and zero outside is called a rectangular window. The Hann window named after Julius von Hann and also known as the Hanning, von Hann and the raised cosine window is defined by

$$w(n) = 0.5 \left(1 - \cos \left(\frac{2\pi n}{N-1} \right) \right)$$

No of filter applied are 12,22,32,42. Maximum accuracy i.e 85% is achieved in 32 no of filters. Therefore, it is proof that too few and too many filter does not give maximum accuracy. Again, keeping 32 no. of filter, author shows that Hanning window gives more accuracy level which is 75% than Rectangular window.

Author of [2] have designed a security system based on speaker identification. MFCC has been used as feature and VQ as feature matching. Variation is applied on no. of Code Book and windowing techniques. Triangular, Rectangular and Hamming windows are applied during feature extraction. Code Book size started from 1 to 64. They use both linear scale and mel scale. 100% accuracy has been achieved using triangular and Hamming

window when no. of code book is 64. Accuracy rate is equal in both linear scale and mel scale. But rectangular window give less accuracy which are 85.7% in linear scale and 95.24% in mel scale. It is proved that triangular window and hamming window are most effective windowing technique for speaker identification.

Author of [3] use Formant as feature which is modeled using GMM-UBM for speaker verification. Acoustic theory tells that the transfer function of energy from the excitation source to the output can be described in terms of the natural frequencies or resonances of the tube. Such resonances are called formants for speech. They represent the frequencies that pass the most acoustic energy from the source to the output [5]. One problem is that formant features like F1, F2, F3, as well as the corresponding bandwidths, are usually not independent from each other [4]

The signals are downsampled to 8 kHz. The LPC analysis is set to find 4 formants. Out of these only first three are considered. The analysis window is Hamming window which length was at 0.049 seconds, the LPC order was 12. Long Term Distribution of LPC formant estimation (LTF) method has been used. LTF analysis can capture anatomical vocal tract characteristics as well as habitual speaker specifics, such as a palatalized setting and other supralaryngeal voice qualities. LPC analysis is set. Authors have found that two formant features gives the worst performance. When they use all three formants, a better performance is achieved. By including the corresponding formant bandwidths, the equal error rate can be reduced additionally. The best performance with an equal error rate of 3% can be observed using all three formant frequencies and the corresponding bandwidths.

Author of [6] use MFCC as feature and Feed Forward Back Propagation Neural Network (FFBNN) for classification for speaker identification. Feature extraction process is accomplished by Continuous Wavelet Transformation (CWT) and MFCC. The CWT scale-determination is a very challenging problem because of its non-stationary nature.

The wavelet transform is a fundamental mathematical tool for multi-resolution or multi-scale decomposition of signals, capable of providing the time and frequency information simultaneously. It involves mapping signals onto time-scale space with superior time-frequency localization as compared to the short-time Fourier transform (STFT), which in some cases cannot show the locations of discontinuity in signals. The continuous wavelet transform (CWT) is based on the correlation of a given continuous-time signal with a set of filters defined by a family of basis wavelet functions [7].

Author has selected Morlet wavelets. The speech signal has been decomposed into CWT sub signal (d1, d2,...,dj). MFCC is designed from these sub signals. It has been observed that the use of CWT improves the recognition rate from 85% to 99.89, which is due to the use of CWT and MFCC. This result proves the system robustness by CWT.

Authors of [8] uses MFCC as feature and Support Vector Machine for speaker verification. Text-dependent speaker verification system is designed by using 20 male and 20 female speakers selected from the Aurora 2 database. All of the test speech signals are noisy-free and are sampled at 8000 Hz with 16-bit resolution. Each test speech signal consists of 2~8 English digital numbers or English alphabets. Performance of the system is reported using the false acceptance rate (FAR), the false rejection rate (FRR), and the equal error rate (EER).

Authors give emphasize on order of MFCC. This paper compared the results obtained on the SVM based speaker verification system with 13 settings of MFCC order, namely $p = 2q$, $q = 1 \sim 13$. An impostor model was trained on all the MFCCs in the impostor data set while the speaker model was built using the corresponding speaker data set. During speaker verification task, a likelihood ratio was computed between the speaker model and the impostor model. The better performance could be obtained when MFCC order $p = 22$. An EER of 0% and average accuracy rate of 95.1% are achieved using the proposed system.

III. Discussion

Both Author of [1] and [2] used MFCC and VQ. In paper [1] authors got maximum accuracy i.e. 85% when 32 no of filters are used without considering the type of window. Accuracy is measure by considering False Acceptance and False Rejection. Author of paper [2] apply the system on 13 male and 8 female speakers. Identification rate was defined as the ratio of the number of speakers identified to the total number of speakers tested. Maximum accuracy i.e. 100% is achieved by using no. of code book is greater than or equal to 16 in both Triangular and Hamming window. It is shown in both the paper that rectangular window has less effective for speaker verification. Author of [6] show that MFCC calculation based on CWT give accuracy rate 99.89%. Author of [8] used MFCC and Support Vector Machine for speaker verification. Effect of order of MFCC is shown here. Average accuracy rate 95.1% was achieved when MFCC order is 22.

IV. Conclusion

Different factors like type of window, no of filter, no of code, order and CWT used during MFCC calculation play a role in speaker verification. From the above discussion, it can be stated that 100% accuracy level can be achieved with feature MFCC and patter matching technique VQ.

References

- [1]. Vibha Tiwari, "MFCC and its applications in speaker recognition", International Journal on Emerging Technologies 1(1): 19-22(2010)
- [2]. Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, "Speaker Identification Using Mel Frequency Cepstral Coefficients", "3rd International Conference on Electrical & Computer Engineering, ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh.
- [3]. Timo Becker¹, Michael Jessen², Catalin Grigoras, "Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models", Interspeech 2008 Special Session: Forensic Speaker Recognition - Traditional and Automatic Approaches. Brisbane, Queensland, Australia. September 2008
- [4]. P. Rose, Forensic Speaker Identification. Taylor & Francis,
- [5]. L. Robiner, B.H.Juang, B.yegnanarayana, "Fundamentals of Speech Recognition", Pearson Education
- [6]. Wael Al-Sawalmeh, Khaled Daqrouq, Omar Daoud, Abdel-Rahman Al-Qawasmi, "Speaker Identification System-based Mel Frequency and Wavelet Transform using Neural Network Classifier", European Journal of Scientific Research ISSN 1450-216X Vol.41 No.4 (2010), pp.515-525 © EuroJournals Publishing, Inc. 2010 <http://www.eurojournals.com/ejsr.htm>
- [7]. A.E. Mahdi, E. Jafer, "Two-Feature Voiced/Unvoiced Classifier Using Wavelet Transform", *The Open Electrical and Electronic Engineering Journal*, 2008, Vol: 2, Page: 8-13
- [8]. Shi-Huang Chen and Yu-Ren Luo, "Speaker Verification Using MFCC and Support Vector Machine", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, March 18 - 20, 2009, Hong Kong