Performance Comparison of MFCC for Text Independent Speaker Identification

Niranjan Samudre
Research Scholar
JJT University, Rajasthan

Abstract— In this paper, a Text-Independent Speaker Identification system is implemented. The Mel Frequency Cepstral coefficients (MFCC's) have been used for feature extraction and Vector Quantization (VQ) technique is used to manipulate the data such that it maintains the most prominent characteristics. The extracted speech features (MFCC's) of a speaker is quantized to a number of centroids using the K-mean algorithm. These centroids constitute the codebook of that speaker. MFCC's are calculated in both training and testing phase. The speaker is identified according to the minimum quantization distance which is calculated between the centroids of each speaker. The performance of the Mel-Frequency Cepstrum Coefficients may be affected by the different factors like number of filters, test shot length, codebook size and the type of window. In this paper, performance comparison of MFCC for the above factors is done to find a best implementation for Text Independent Speaker Identification.

Key words: Speaker Identification, MFCC, Text Independent Speaker Identification

I. INTRODUCTION

Modern-day security systems are wide-ranging and usually have multiple layers to get through. Besides the standard locks and deadbolts and alarm systems, there are very complex methods to protect the important material. Many of these are methods that permit or restrict a specific individual to access the information, e.g. a computer system with fingerprint recognition, individual's eye pattern recognition, or voice recognition.

Speaker recognition has been an interesting and challenging research field for the last decades, which still yields a number of unsolved problems. Speaker recognition is basically divided into speaker identification and speaker verification [1]. Verification is the task of automatically determining if a person really is the person he or she claims to be. Speaker identification consists of mapping a speech signal from an unknown speaker to a database of known speakers which the system can recognize. Speaker Identification systems can be subdivided into text-dependent and text-independent methods. Text-dependent systems require the speaker to utter a specific word (pin-code, password etc.), while a text-independent method catches the characteristics of the speech irrespective of the text spoken [14][23].

The focus of this paper is Text-Independent Speaker Identification, meaning the system can identify the speaker regardless of what is being said. This technique consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker- specific features. Using training data these features are clustered to form a speaker-specific codebook. In the testing stage, the test data is compared to the codebook of each reference speaker and a measure of the difference is used to make the identification decision. The feature extraction is done using Mel Frequency Cepstral Coefficients (MFCC) [1].

The performance of the Mel-Frequency Cepstrum Coefficients is studied for the different factors like number of filters, test shot length, codebook size and the type of window to find a best implementation for Text Independent Speaker Identification.

II. LITERATURE REVIEW

The speaker identification problem has been addressed in the literature by representing the speech signal using different features, calculated in the frequency and in the time domain.

Today, to represent speech signal various important features like energy, pitch frequency, format frequency [7, 8], linear prediction coefficients (LPC), linear prediction cepstral coefficients (LPCC), Mel-Frequency cepstral coefficients (MFCC) are used [9,10].

MFCC's has been widely used in many audio processing task, *e.g.*, language identification [9], speech emotion classification [10], and speaker identification [11]. MFCC's provided a compact representation of the spectral envelope of a frame of speech. Among all cepstral analysis, MFCC's are proved to be very significant in speaker identification and verification system [7]. MFCC was found to be the optimum feature extraction technique by Douglas and Richard in [6]. In [24], MFCC and LPCC were the techniques used for extracting the features. On their analysis they found good performance by MFCC for extracting the features. In [5], various feature extraction methods like MFCC, LPCC and MMFCC were compared for speech enhancement in noisy environment and MFCC was selected as the optimum feature extraction technique.

Other authors have proposed the use of acoustic features directly obtainable from the time domain, such as pitch, speech rate, voice quality and temporal variation of the audio signal.

III. SPEAKER IDENTIFICATION

Speaker identification is comparing a speech signal from an unknown speaker to a database of known speakers. The system has been trained with a number of speakers which the system can recognize. Speaker identification can be further divided into two branches. Open-set speaker identification decides to whom of the registered speakers' unknown speech sample belongs or makes a conclusion that the speech sample is unknown. In this work, we deal with the closed-set speaker identification, which is a decision making process of whom of the registered speakers is most likely the author of the unknown speech sample. Depending on the algorithm used for the identification, the task can also be divided into text-dependent and text-independent identification. The difference is that in the first case the system knows the text spoken by the person while in the second case the system must be able to recognize the speaker from any text [6].

The process of speaker identification is divided into two main phases. During the first phase, speaker enrollment, speech samples are collected from the speakers, and they are used to train their models. The collection of enrolled models is also called a speaker database. In the second phase, identification phase, a test sample from an unknown speaker is compared against the speaker database. Both phases include the same first step, feature extraction, which is used to extract speaker dependent characteristics from speech. The main purpose of this step is to reduce the amount of data while retaining speaker discriminative information. In the enrollment phase, these features are modeled and stored in the speaker database. This process is represented in Figure 1.

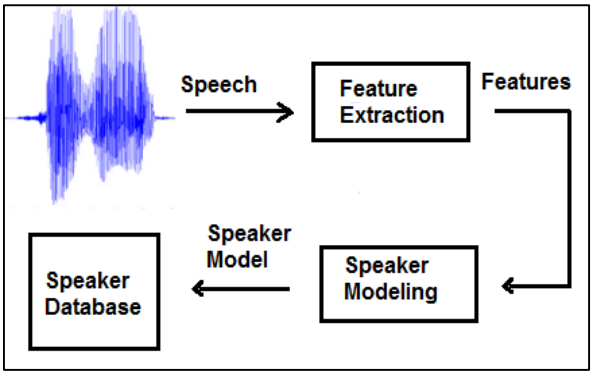

Fig. 1: Enrollment Phase

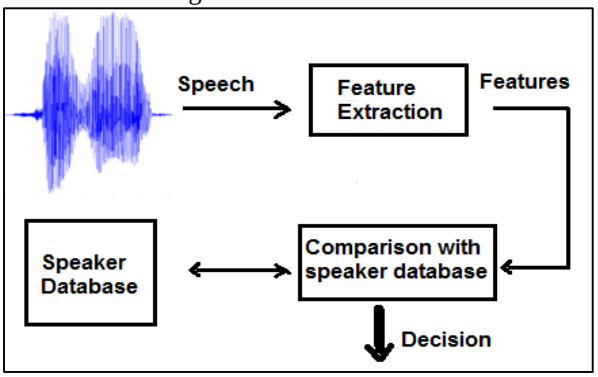

Fig. 2: Identification Phase

In the identification step, the extracted features are compared against the models stored in the speaker database. Based on these comparisons the final decision about speaker identity is made. This process is represented in Figure 2.

A. Feature Extraction

The speech signal contains different kind of information about speaker. This includes "high-level" properties such as dialect, context, speaking style, emotional state of speaker and many others [14]. More useful approach is based on the "low-level" properties of the speech signal such as pitch, intensity, formant frequencies and their bandwidths, spectral correlations, short-time spectrum and others [2].

From the automatic speaker identification point of view, it is useful to think about speech signal as a sequence of features that characterize both the speaker as well as the speech. It is an important step in identification process to extract sufficient information for good discrimination in a form and size which is responsible for effective modeling [4]. The amount of data, generated during the speech production, is quite large while the essential characteristics of the speech process change relatively slowly and therefore, they require less data. According to these matters feature extraction is a process of reducing data while retaining speaker discriminative information [3, 4].

The speech wave is usually analyzed based on spectral features. There are two reasons for it. First is that the speech wave is reproducible by summing the sinusoidal waves with slowly changing amplitudes and phases. Second is that the critical features for perceiving speech by humans ear are mainly included in the magnitude information and the phase information is not usually playing a key role [15]. In this paper Mel Frequency Cepstral Coefficients as features for the classification problem are used.

B. Framing and Windowing

The speech signal is slowly varying over time, so when the signal is examined over a short period of time (5-100msec), the signal is fairly stationary. Therefore speech signals are often analyzed in short time segment, which is referred to as short-time spectral analysis [3]. It works as follows: predefined length window (usually 20-30 milliseconds) is moved along the signal with an overlapping (usually 30-50% of the window length) between the adjacent frames as shown in figure 3. Overlapping is needed to avoid losing of information. Parts of the signal formed in such a way are called frames. In order to prevent an abrupt change at the end points of the frame, it is multiplied by a window function. The operation of dividing signal into short intervals is called windowing and such segments are called windowed frames. The most popular window function used in speaker identification is Hamming window function, which is described by the following equation:

$$h(n) = 0.54 - 0.46 \cos(2\pi n / N - 1),\ 0 \leq n \leq N - 1 \qquad (1)$$

Where *N* is the size of the window or frame. A set of features extracted from one frame is called feature vector.

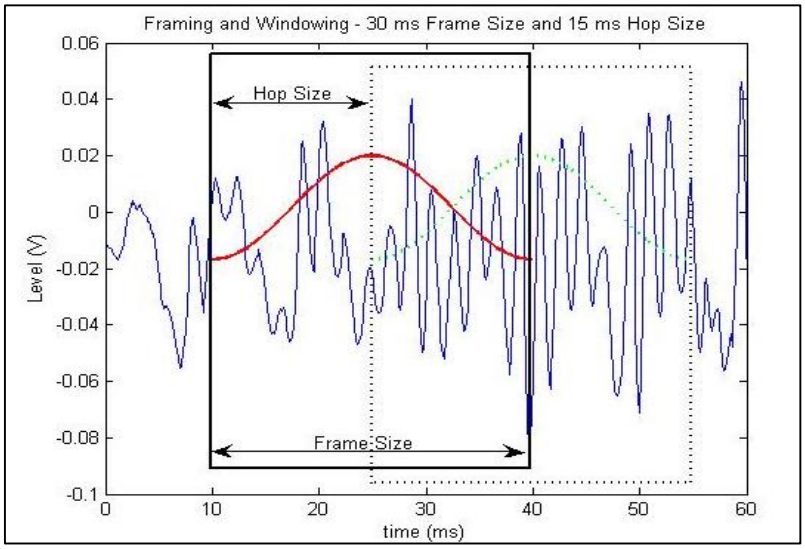

Fig. 3: Framing and Windowing

C. Mel-Frequency Cepstrum Coefficients

Mel-frequency cepstrum coefficients (MFCC) are well known features used to describe speech signal. They are based on the known evidence that the information carried by low-frequency components of the speech signal is phonetically more important for humans [3]. MFCC computing is based on the short-term analysis, and thus from each frame a MFCC vector is computed. MFCC extraction is similar to the cepstrum calculation except that the frequency axis is warped according to the mel-scale. Summing up, the process of extracting MFCC from continuous speech is illustrated in Figure 4.

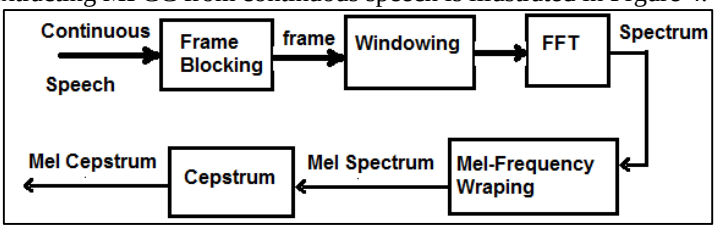

Fig. 4: MFCC Block Diagram

The speech signal consists of tones with different frequencies. For each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on the 'Mel' scale. The *mel-frequency* scale is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz [17].

We can use the following formula to compute the mels for a given frequency *f* in Hz [25]:

$$mel(f) = 2595 * \log10(1 + f/700) \qquad (2)$$

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired mel frequency component. The filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval.

The DCT implements the same function as the FFT more efficiently by taking advantage of the redundancy in a real signal. The DCT is more efficient computationally.

The MFCCs may be calculated as:

$$(C_n)^{\sim} = \sum_{k=1}^{Kn} (log\ S_k)[n(k - \tfrac{1}{2})\pi/K \qquad (3)$$

Where $n = 1, 2, \ldots, K$

D. Vector Quantization

Speaker identification systems are inherent of a database, which stores information used to compare the test speaker against a set of trained speaker voices. Practically feature extraction with a very large data cannot be achieved. The number of feature vectors would be so large that storing and accessing this information using current technology would be impractical [18].

Vector Quantization (VQ) is a quantization technique used to compress the information and manipulate the data in such a way to maintain the most prominent characteristics. VQ is used in many applications such as data compression (i.e. image and voice compression), voice recognition, etc. Given the extracted feature vectors (known as codeword) from each speaker, each codeword is used to construct a codebook. This process is applied to every single speaker to be trained into the system. Although numerous VQ algorithms exist, Linde-Buzo-Gray algorithm is chosen, since it is the easiest to implement [19].

E. Decision

The decision making logic is handled by a concept known as the threshold. The threshold determines the acceptable boundaries giving the final answer. The system will only result in a solution if the following criteria are met.

1) The system has found the lowest Euclidean Distance between the codebook tested and the various trained codebooks.
2) The distance calculated falls below a pre-defined threshold of acceptance.

Both requirements must be satisfied in order for the system to produce a result, otherwise the voice signal in test will be given as an "unknown speaker".

The formula used to calculate the Euclidean distance [16] can be defined as following:

The Euclidean distance between two points *P = (p1, p2…, pn) and Q = (q1, q2..., qn),*

$$\sqrt{(p1 - q1)^2 + (p2 - q2)^2 + \cdots + (pn - qn)^2} = \sum_{i=1}^{n} (pi - qi)^2 \qquad (4)$$

The speaker with the lowest distortion distance is chosen to be identified as the unknown person.

IV. IMPLEMENTATION

A. Data Description

In order to evaluate the performance of the system as a real time application database is formed which consists of 8 speakers, 4 female and 4 male, their ages are about 20 years old [16]. Training set consists of 30 seconds speech utterance, where the testing sessions were done using three different test shot lengths.

B. Results

The following figures supports that our system identifies the speaker on the basis of the speaker voice characteristics regardless the uttered text.

1) Four different utterances for the same speaker

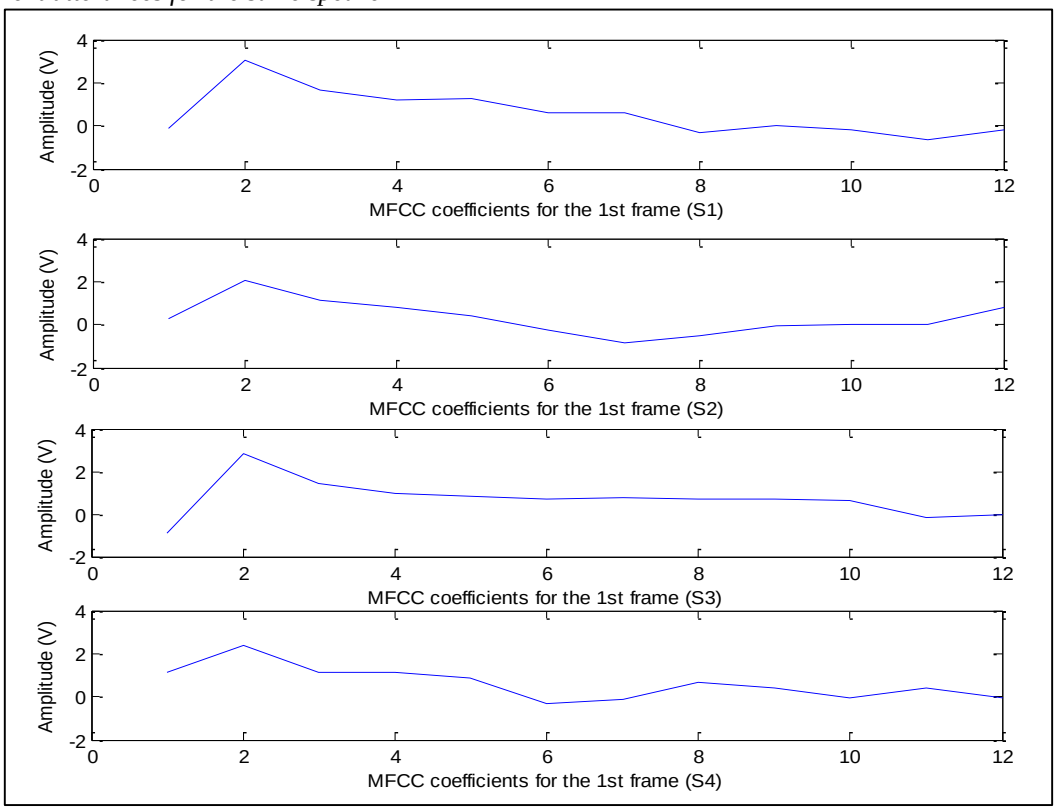

Fig. 5: MFCC coefficients for different words from the same speaker.

It can be seen in the figure 5 that although the speaker said different utterances; but the shape of the envelope containing the coefficients in general looks the same. This supports the idea that our system recognize the speaker regardless what he/she said.

2) The same utterance by four different speakers

It can be seen in the figure 6 that the shape is different from one speaker to another, its unique for each speaker.

3) The effect of the codebook size on the identification rate

The identification rate for the number of centroids is given the table 1.

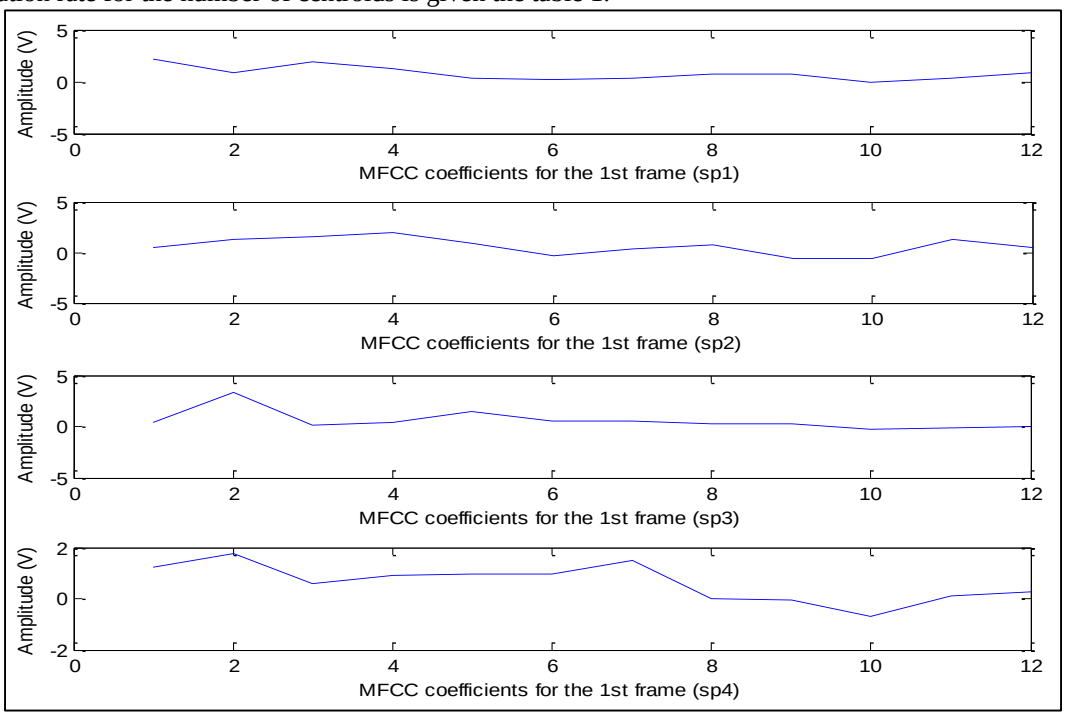

Fig. 6: MFCC coefficients for same words from the different speakers.

Number of centroids (C)	Identification rate (%)
2	84.375
8	95
16	98.75
64	98.75

Table 1: Codebook Size V/S. Identification Rate.

The effect of the codebook size on the identification rate can be explained as following; as we increase the number of centroids, this means that the number of clusters will increase, so each cluster will contain less number of feature vectors. As a result, the average vector for each cluster will be with less amplitude's values. So the difference terms in the Euclidean distance will have less value. Therefore, the distance will decrease and the Identification rate will decrease.

4) The number of the MFCC coefficients.

Increasing the number of mel frequency cepstral coefficients results in improving the identification rate on the expense of the computational time. MFC coefficients are typically in the range (12-15).

No. of MFC coefficients	Identification Rate (%)
5	76
12	91
20	91

Table 2: Identification Rate V/S The Number Of The Mfc Coefficients

5) The number of filter-banks

To study the effect of changing the number of filters in the filter-bank on the identification rate, the tests were performed using test speakers for different values of the filter-banks and the identification rate for each value is calculated. The results are given in the table 3.

Number of Filter-banks	Identification Rate (%)
12	15.75
16	60.75
20	85.62
30	96.75

Table 3: Number Of Filter-Banks V/S. Identification Rate Using A Codebook Size Of 64 And 12 Mfcc.

It is obvious that number of the filter-banks plays a major role for the purpose of improving the identification accuracy.

6) The performance of the system on different test shot lengths.

To study the performance of different test shot lengths, three tests were conducted using eight test speakers uttering the test speech sample with three different lengths. The results are shown in the table 4.

Test speech length	Identification rate (%)
5 sec	75
15 sec	85
25 sec	95

Table 4: Identification Rate for Different Test Shot Lengths.

It can be concluded that the Identification rate increases with large test shot length.

7) Effect of variation in type of window using 32 filters

Considering 32 filters as a standard number of filters we have changed the window type. In this experiment two windows are used *viz*. Hanning Window and Rectangular window. Results show that efficiency is maximum while using hanning window.

a) Hanning window

Speaker	No. of Attempts	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	0
S3	4	0	3
S4	4	0	0
S5	4	0	2
Total	**20**	**0**	**5**

Table 5: Hanning window

Threshold value of distance = 150, Efficiency = 75%

b) Rectangular window

Speaker	No. of Attempts	False Acceptance	False Rejection
S1	4	0	0
S2	4	0	2
S3	4	0	3
S4	4	0	0
S5	4	0	4
Total	**20**	**0**	**9**

Table 6: Rectangular window

Threshold value of distance = 150, Efficiency = 55%

V. CONCLUSION

The Text-Independent Speaker Identification system is implemented. The feature extraction is done using Mel Frequency Cepstral Coefficients (MFCC) and the speakers were modeled using Vector Quantization technique. Using the extracted features, a codebook from each speaker was build. Clustering of the feature vectors is done using the K-means algorithm. Codebook from all the speakers was collected in a database. A distortion measure based on minimizing the Euclidean distance was used when matching the unknown speaker with the speaker database.

The performance comparison of MFCC for the different factors like number of filters, test shot length, codebook size and the type of window is done to find a best implementation for text independent speaker identification. The study shows that as the number of centroids increases, the identification rate of the system increases. Also, the number of centroids has to be increased as the number of speaker's increases. The study shows that as the number of filters in the filter-bank increases, the identification rate increases. Results also showed that reducing the test shot lengths reduced the identification accuracy. In order to obtain satisfactory result for real time application, the test data usually needs to be more than ten seconds long.

REFERENCES

[1] L.R. Rabiner and R.W. Schafer, 'Digital Processing of Speech Signals', New Delhi: Prentice Hall of India. 2006.
[2] B. S. Atal, "Automatic Recognition of Speakers from their Voices", Proceedings of the IEEE, vol. 64, 1976, pp 460 – 475.
[3] J. R. Deller, J. H. L. Hansen, J. G. Proakis, 'Discrete-Time Processing of Speech Signals', Piscataway (N.J.), IEEE Press, 2000.
[4] H. Gish and M. Schmidt, "Text Independent Speaker Identification", IEEE Signal Processing Magazine, Vol. 11, No. 4, 1994, pp. 18-32.
[5] Md. Rabiullslam, Md. Fayzur Rahmant, Muhammad Abdul Goffar Khant, "Improvement of Speech Enhancement Techniques for Robust Speaker Identification in Noise", Proceedings of International Conference on Computer and Information Technology, December 2009.
[6] D. A. Reynolds, R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, 1995, pp. 72 -83.
[7] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Tranaction on Acousiic., Speech, and Signal Processing. 29(2): 254-272, 1981

[8] S. Slomka and S. Sridharan, "Automatic Gender Identification Optimized for Language Independence", Proceedings of IEEE TENCON'97, pp. 145-148, Dec. 1997.
[9] E. S. Parris and M. J. Carey, Language Independent Gender Identification, ICASSP, pp 685-688, 1996.
[10] Ting, H, Yingchun, Zhaohui, W., Combinning MFCC and Pitch to Enhance the Performance of the Gender Recognition, IEEE, 2006
[11] U. Bhattacharjee and K. Sarmah, "Language Identification System using MFCC and Prosodic Features", in Intelligent Systems and Signal Processing (ISSP), pp. 194-197, 2013.
[12] Z. M. Dan and F. S. Monica, "A Study about MFCC Relevance in Emotion Classification for Srol Database" in Electrical and Electronics Engineering (ISEEE), pp. 1-4, 2013.
[13] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task". in Proceedings of 10th International Conference on Speech and Computer. Vol. 1, pp. 191-194, 2005.
[14] J. M. Naik, "Speaker Verification: A Tutorial", IEEE Communications Magazine, January 1990, pp.42-48.
[15] S. Furui, Digital Speech Processing, Synthesis and Recognition, New York, Marcel Dekker, 2001.
[16] N. A. Samudre," Text Independent Speaker Identification using Vector Quantization", International Journal of Engineering Research & Technology (IJERT), Volume 2, Issue 8, August 2013, ISSN: 2278 – 0181, ESRSA Publication.
[17] Seddik, H.; Rahmouni, A.; Sayadi, M.; "Text Independent Speaker Recognition using the Mel Frequency Cepstral Coefficients and a Neural Network Classifier" First International Symposium on Control, Communications and Signal Processing, Proceedings of IEEE 2004 Page(s): 631-634.
[18] C.-H. Lee, F.K. Soong, K.K. Paliwal: "Automatic Speech and Speaker Recognition" - advanced topics. Kluwer Academic Publishers, pp. 42-44, Norwell, Massachusetts, USA, 1996
[19] S. Sookpotharom, S. Manas "Codebook Design Algorithm for Classified Vector Quantization" Bangkok University, Pathumtani, Thailand pp. 751-753 2002
[20] M.Brooks, Voicebox: Speech Processing Toolbox for MATLAB, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox
[21] J. R. Deller, J. G. Proakis and J. H. L. Hansen, Discrete-time Pro-cessing of Speech Signals, Prentice Hall, New Jersey, 1993.
[22] L. Feng, Speaker Recognition, Master's thesis, Technical University of Denmark, Informatics and Mathematical Modelling, 2004, ISSN: 1601- 233X.
[23] J. P. Campbell JR, Speaker Recognition: A Tutorial, in Proceedings of the IEEE, vol. 85 no. 9, 1997.
[24] E. Karpov, Real-Time Speaker Identifcation, Master's thesis, University of Joensuu Department of Computer Science, 2003.
[25] V. Tiwari, "MFCC and its Applications in Speaker Recognition", International Journal on Emerging Technologies, volume 1 (1), 2010, pp 19-22.