Performance of Complementary Features for Robust Speaker Identification

Sharada V. Chougule
Assistant Professor, Finolex Academy of Management &Technology, Ratnagiri, Maharashtra

Mahesh S. Chavan
Professor, KIT's College of Engineering, Kolhapur, Maharashtra

ABSTRACT

This paper considers the problem of acoustic mismatch caused by use of different sensors, in digital gazettes and hand-held devices. In this paper, two complementary features derived from conventional cepstral features are proposed, namely linear/mel spectral subband features (L/M-SSC) and log filter bank energy features (LFBE). The performance of these complementary features is compared with conventional features in acoustic mismatch conditions. To investigate the performance of features alone, all processing and classification steps are kept constant to allow a controlled comparison. A multi-variability speech database (IITG-MV) with acoustic mismatch (different microphones) is used for experimental evaluation. It is observed that all these features shows almost equal performance for text independent speaker identification in same acoustic condition. Whereas in mismatch condition, spectral subband centroids (L/M-SSC) features proved to be robust than other features when used alone. Further, use of dynamic features along with channel and noise compensation enhances the percentage identification rate of the system for all cases of acoustic mismatch, with spectral subband centroid features showing comparable performance to that of conventional features.

Keywords

MFCC, LFCC, Linear/Mel scale spectral subband centroids (L/M-SSC), Log filter bank energy (LFBE)

1. INTRODUCTION

Speech is a natural and the main way of communication of human being to convey message between each other. Along with the conveying the message, it also convey other form of information indirectly such as identity of a person (type of voice), the language spoken, emotion, health and also in some case social (educational) background of the person. To recognize a person from his/her voice alone, our brain use several different types (or levels) of information and clues present in the speech. Automatic speaker recognition is a similar task performed by machines with speech signal as a input and identity of a particular person as a output. Depending on the type of end decision, speaker recognition is classified as *speaker identification* and *speaker verification*. In speaker verification the goal is to verify the identity claim of a speaker and either accept or reject the claimant. Speaker identification determines which voice in a known group of voices best matches the speaker [1]. The system can be further classified as text dependent and text independent based on nature of speech data (same or different) used for training and testing sessions. A closed-set text independent speaker identification (TISI) system identifies a person from a set of known samples of speech data, whereas in open set, the input speech sample need not be previously stored in the system's database and the end result could be of the form "none of the above".

This paper investigates the performance of complementary features for acoustic mismatch. Section 2 discusses the motivation and literature survey. Conventional and complementary feature extraction techniques are discussed in Section 3, with experimental analysis in Section 4 and conclusion in Section 5.

2. MOTIVATION AND LITURATURE SURVEY

Using the speaker recognition systems in variety of applications (e.g. person verification for banking transaction over phone, identifying a specific speaker in multi-media recordings, speaker recognition for home and industry service robots, access to systems and servers in military and security related applications) faces the problem of acoustic mismatch caused by use of different types of devices (sensors) used to collect and test the system. With ubiquitous portable devices and ever increasing use of multimedia web portals, market for speaker recognition systems is growing exponentially. Mismatch between training and testing sessions is one of the main impediments in achieving the precise performance in real world applications. With the satisfactory performance of speaker recognition system in laboratory conditions, the research is now inclined to enhance the robustness of speaker recognition system in mismatch condition. This mismatch comes from variety of factors (except interspeaker variability) such as change of surrounding environment, transmission channel, using different handsets or microphones, or may be due to psychological and pathological state of the speaker and the linguistic contents [1].

In view of the challenge of handling the mismatch in practical applications of speaker recognition system, this work is motivated by the study of some conventional and some complementary features for speaker recognition. The objective behind this work is twofold. First is to study the various features and their characteristics representing speaker specific parameters (clues) and then analyze the performance of these features for mismatch condition observed in practical applications for text independent speaker identification.

Feature extraction and model formation are the two basic stages in building the speaker recognition system. Robustness to mismatch can be obtained by making the features to be robust or to have a model which can sustain its performance in variety of situations. In last decade, much of the research is dedicated towards development of speaker models dealing with the variety of session mismatch. Vector quantization is one of the simplest model with the base of data compression, used for text independent speaker recognition. Its main features are high computational speed and light weight practical implementation [2] and requires comparative less training data. Gaussian mixture model (GMM)[3] is conventionally used probabilistic modeling technique for text independent speaker recognition. Its advantage is the effectiveness and scalability in modeling the spectral distribution of speech, whereas disadvantage is the requirement of sufficient data to model the speaker. This drawback is overcome using a universal background model (UBM) to form a speaker independent model by pooling the speech data from large number of speaker, which act as a speaker model [4],[5]. Auto-associative neural network (AANN) is an alternative to GMM developed for pattern recognition task studied for speaker recognition in [6]. Support vector machine (SVM) is a powerful discriminative classifier based on designing a proper kernel metric which separates the target speakers from the UBM speakers by establishing a hyperplane by training in one-versus-all manner[7], studied with Bhattacharyya based distance in [8]. GSV-SVM and MLLR-SVM approaches are used to build acoustic models trained using speaker adaptive training(SAT) shown to outperform GMM-UBM system using hybrid factor analysis and SVM alone [9].

Feature extraction is considered as a heart of any speech as well as speaker recognition systems. However, its role in the speech system is completely contradictory to that of the speaker recognition and vice-versa. The goal of feature extraction stage in text independent speaker recognition is to derive a set of features relevant to the speaker irrespective of the spoken words and linguistic contents.

Although no specific feature can completely characterize a particular speaker's voice, speaker specific attributes are always present in some form in one's speech. The goal of feature extraction stage is to extort these parameters unique to the speaker and eliminate all other. The speaker specific information can be categorized broadly into two categories: low level features and high level features [2]. Low level features describes the characteristics of human vocal tract (called physical characteristics), whereas high level features represents the behavioral features of the speaker such as conversational patterns, prosody, idiolect etc. Low level features are also called segmental features, as these are computed over short time 20-30 ms , whereas high level features are commonly known as supra-segmental features which are observed over a longer time interval greater than few seconds.

Cepstral features derived from short time spectrum of speech signal are very useful and popular for audio processing in clean environment and matched conditions. Mel Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Cepstral Coefficients (LPCCs) are well known examples of the same. However sensitivity to noise and mismatch conditions is the main disadvantage of cepstral features. There has been increased efforts in recent time to improve the performance (accuracy) of speaker recognition in these uncontrolled situations. In the next paragraph we will summarize the work done at feature level to enhance the robustness of the system.

Speaker Identification with robust front end processing using unsupervised speech activity detection (SAD), combined with perceptual spectral flux is proposed by Sadjadi et.al., using mean Hilbert envelop coefficients (MHEC) and found to outperform in extremely degraded communication channel [10]. Conventional MFCC use short time (20-30 ms) frames to extract features vectors. Instead, use of longer frame length and window shown improved results in the presence of white Gaussian noise and mid-range SNR [11]. Authors in [12], make use of Gammatone Frequency Cepstral Coefficients (GFCC) features and compared it's performance with MFCCs for noise robustness in speaker identification. Effect of variation in various parameters is analyzed for various noises of different SNRs. The study concludes that GFCC features are more robust to noise than MFCC features. Modulation features of medium duration subband speech amplitudes (MMeDuSA) was proposed [13] for noise robust speaker recognition and compared with MFCC, PNCC, MHEC and MDMC respectively.

3. FEATURE EXTRACTION

Any speaker identification system involves feature extraction (along with some front end processing) as the initial step. Front end module transforms the speech signal in the form suitable for further processing. 'Features' are nothing but a compact and appropriate representation of speech signal, which is more stable and discriminative than the original speech. For speaker recognition, the extracted features should carry the characteristics of an individual voice ignoring linguistic and other contents. Also for a robust system, it is required that the features should also be robust in practical conditions (such as noise and various types of mismatch. We now discuss the steps in computation of the features used in this study.

3.1 Mel Frequency Cepstral Coefficients (MFCCs) Features

Davis and Mermelstein in 1980 invented MFCCs to extract phonetic features for word recognition [14]. It is conventionally used in speaker recognition to reflect human vocal tract characteristics depending on its shape and length. Ease of computation and reliability in clean environment are the main attributes of the same. A short term spectrum of the speech signal is obtained by dividing the entire speech into small number of frames (typical size 10-25 ms) and windowing the same with overlap (5-10 ms typically) to avoid spectral leakage through direct framing. Also response of each frequency is completely uncorrelated using windowing function. The FFT spectrum obtained is passed through a set of filter-bank (called mel-scale filter bank) where the filters are spaced linearly at low frequency (below 1 kHz) and logarithmically at high frequencies (above 1 kHz) to mimic the known variations of ear's critical bandwidths with frequency. The mel scale is given by [14]:

$$fmel = 2595 * \log(1 + \frac{f}{700}) \qquad (1)$$

Energies from output of each filter is then computed at time instance 't' are given by:

$$e[j][t] = \sum_{k=0}^{N-1} H_j(k) * |\tilde{S}_t(k)|^2) \qquad (2)$$

for j=1,.....,P, where $H_j(k)$ are P triangular filters and $|\tilde{S}_t(k)|^2$ represents the signal power spectrum. respectively.

Taking logarithm of these energies compresses dynamic range of values and makes frequency estimates less sensitive to slight variations in input. Finally performing inverse DFT called discrete cosine transform (DCT) on log spectrum produces highly uncorrelated features representing the vocal tract characteristics, called mel frequency cepstral coefficients (MFCCs), given by:

$$MFCC[i][t] = \sqrt{\frac{2}{P}} \sum_{j=1}^{P} \left\{ \log e[j][t] * \cos(\frac{\pi i}{P}(j-0.5)) \right\} \quad (3)$$

3.2 Linear Coefficient Cepstral Coefficients (LFCC) Features

The computation of LFCC features is similar to MFCC discussed above, except the nature of filter bank used weight the FFT spectrum. Here 26 linearly spaced, overlapped filters are used (instead of mel-warped filters), which gives equal weight to all the frequencies through out the spectrum. Log and DCT is further applied to separate and de-correlate the complex features.

3.3 Spectral Subband Centroids (SSC) Features

MFCC features captures the shape of speech spectral envelop based on subband magnitude spectrum using a mel scale filter bank. The smoothed power spectrum may cause loss of some information in the presence of noise. Spectral Subband Centroids [15], [16] are a set of centroids confined to be within each spectral subband. It is an alternative feature to cepstral features (like MFCC s and LPCCs). SSC provide different information than MFCC in the sense that, it computes the peaks in the power spectrum in each subband, which are less affected by noise than the weighted amplitude of power spectrum in case of MFCC. It has been shown in [1] that SSCs are closely related to position of spectral peaks (formants) of speech sounds and proved to be robust in the presence of white and babbling noise [16].

For computation of SSCs, the entire frequency band (0 to Fs/2) is divided into M number of subbands, where Fs is the sampling frequency of the speech signal. SSCs are found by applying filter bank to the power spectrum of the signal and then calculating first moment (centroid) of each subband [16]. With lm and hm are the lower and upper edges of the subband, the m^{th} subband centroid is defined as in [15]:

$$Cm = \frac{\int_{lm}^{hm} f * w_m(f) P^{\gamma}(f) df}{\int_{lm}^{hm} w_m(f) P^{\gamma}(f) df} \quad (4)$$

where $w_m(f)$ is the shape of filter and $p^{\gamma}(f)$ be the power spectrum at the location 'f' raised to the power of γ, which is a constant used for controlling the power range of the power spectrum.

The parameter $w_m(f)P^{\gamma}(f)$ decides where the centroid should be. In our case, $w_m(f)$ is triangular shaped with power parameter set to one as it is not motivated by any psychological aspect of hearing .

3.4 Log Filter Bank Energy (LFBE) Features

Cepstral features represents the smooth envelop of short-time frame of a speech signal. The conventional mel cepstrum comes from log energies (LFBE) ,S(k) for k=1,2,...P of a set of P mel spaced filters. Further a compact and quasi-correlated representation of feature vectors are obtained with the use of discrete cosine transform (DCT) in mel-cepstrum. In order to study the effect of log filter bank energies on speaker identification, the transformation of the sequence S(k) in cepstral domain is avoided by filtering that sequence. The resultant features are called as LFBE features given with reference to equation (2) as:

$$lfbe[j][t] = \log(\sum_{k=0}^{N-1} H_j(k) * |\tilde{S}_t(k)|^2)) \quad (5)$$

4. PERFORMANCE EVALUATION

4.1 Baseline System

Short term spectral analysis is performed on input speech with a window size of 20 ms and frame overlap of 10 ms. A set of feature vectors are carried out from each frame using the feature extraction techniques discusses in section 3. Before actual feature extraction speech signal is pre-emphasized using a simple first order IIR-HP filter of factor of 0.97. Pre-emphasis is requires to avoid *spectral tilt* which is caused by the nature of glottal pulse. Energy in high frequency speech signals is boosted by pre-emphasis which gives more information to acoustic model.

Each of the feature uses 26 filters either non-uniformly scaled (mel-scale) or uniformly scaled (linear). In case of SSC features we distinguish them by naming as M-SSC/L-SSC and for log filter bank energy features referred as M-LFBE/L-LFBE (where M-mel scale, L-Linear scale). The extracted features are processed further to form the model of each speaker. As the goal is to investigate the performance of features alone, all processing and classification steps are kept constant to allow a controlled comparison.

A closed set text independent speaker identification (TISI) system is build with vector quantization (VQ) technique for model formation. The reason for using vector quantization is that it is simple to implement and less amount of speech data is required for training the system. Also as proved in [17], it yields almost equally good performance to that of baseline GMM with maximum likelihood training. In training phase a speaker model (called as codebook) is created from the speech samples of each speaker (N number of speakers), which are stored in the database. In testing (identification) phase, speech data from the input speaker (out of these N speakers for a closed set TISI) is analyzed and compared with the stored database for the best matching model based on distance measure algorithm. A match score is assigned to every speaker and a speaker that yields smallest distortion (in terms of distance) is identified as the best match.

4.2 Database

To evaluate the performance of TISI system in acoustic mismatch, a multi-variability speech database (from EMST Lab, IITG) [18], collected across different sensors is used.

Table 1. Description of Database

Speech Material	Description
Number of speakers	100
Language	English
Speaking style	Read speech
Training duration	Apprx. 30 msec
Testing duration	Apprx. 10 msec
Sensors used	Headset, Table PC (TPC), Digital Voice Recorder(DVR), Mobile Phone

The details of the speech material used for training and testing is given in Table 1. Five different sensors are used to collect the speech data, each with different sampling frequency. An acoustic mismatch is caused due to this, which may degrade the quality of input speech. Fig.1 shows the speech file and spectrogram of same speaker obtained with four different sensors (out of five). As observed, the speech signal from headset microphone is most clear, whereas there is a lot of distortion in the speech signal obtained from digital voice recorder (DVR).

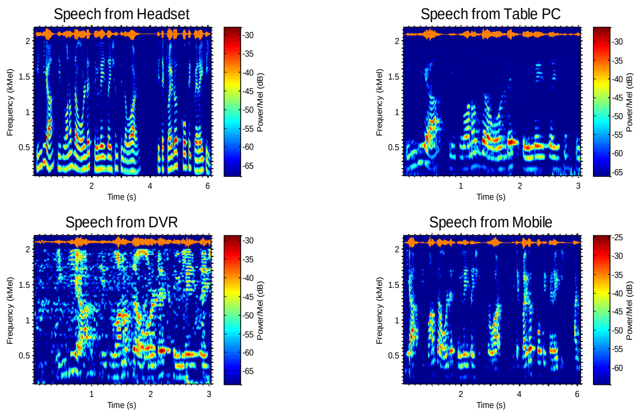

Fig.1 Spectrogram of the same speaker's speech recorded with different microphones

4.3 Results and Discussion

Any cause of distortion or mismatch in speech data can affect the front end processing, and in turn the performance of speaker recognition system in general. The reason for distortion in this case is the type of device used to collect the data which results in acoustic mismatch. As observed from plots in Fig. 2, in matched condition (same sensor for training and testing), MFCC, LFCC and LFBE features show 100 % accuracy, whereas there is a drastic drop in correct identification in case of mismatch. Also it is observed that spectral subband centroid features, with mel scale filter bank proves to be more robust to mismatch in its original form (baseline features).

In order to improve the performance of baseline features, the well known delta and delta-delta features are appended to spectral features. These features represents the dynamic information of the speech spectrum The time derivative of baseline features is estimated using differentiation (called velocity coefficients) and are appended to spectral features. Further the derivative of delta features gives delta-delta (or acceleration) coefficients. Using these dynamic features, the dimension of baseline features is increase by 3*x if x is original feature dimension. The dynamic features captures the time varying information in the speech spectrum which was suppressed by the mismatch.

From Fig. 3, it may be observed that the use of higher dimensional dynamic features improves the identification accuracy of all features for mismatch condition except for the case of test data with digital voice recorder (DVR), the reason is obvious there is a large degradation in the from this sensor.

In acoustic mismatch, some channel noise and additive (unknown /environmental) noise may get introduced due to the placement of microphone (near the lips or on the table) or its type (directional or omnidirectional).

To reduce the time varying distortions introduced due to transmission channel and recording device cepstral mean normalization (CMN) and cepstral variance normalization (CVN) is performed. To overcome the effect of these two, we further modify the features by normalizing the spectral features using mean and variance normalization. Normalizing the variance of cepstral coefficients, helps to improve recognition in adverse conditions [19].We refer the two method together as cepstral mean and variance normalization (CMVN). Fig. 4 shows the increased identification rate for all the features explicitly in case of Headset-Digital voice recorder (H-DVR).

Spectral subtraction (SS) [20] is one of the earliest approach to noise compensation and speech enhancement, used for the suppression of additive noise from the corrupt signal. It is based on method of subtracting the noise estimate (magnitude) from the corrupt spectrum assuming noise to be stationary. So, the normalized dynamic cepstral feature vectors are modified further with spectral subtraction, the result of which is shown in Fig. 5. It is clear from the plots that percentage of correct identification is improved to almost satisfactory level (between 95-100%) with the use of this noise compensation technique for the first four set of features. Log filter bank energy features found to be less robust in acoustic mismatch.

5. CONCLUSION

This paper discusses the performance of conventional and complementary features for acoustic mismatch in text independent speaker identification. The acoustic mismatch is observed due to use of different devices to collect the speech data. Of the various features studied, Spectral Subband Centroid (L/M-SSC) is found to be more robust in mismatch condition as a baseline feature, when used alone. The reason could be the spectral peaks as a feature, which get less affected by noise and distortion. The cepstral features (MFCCs and LFCCs) are found to be much sensitive to acoustic mismatch. However, when modified with dynamic features along with noise and channel compensation techniques, showed improved performance to a satisfactory level. As computation and complexity of these complementary features is less, these features may found useful for other mismatch conditions. The further work will be done to study the performance of these complementary features in various mismatches that may occur in real world.

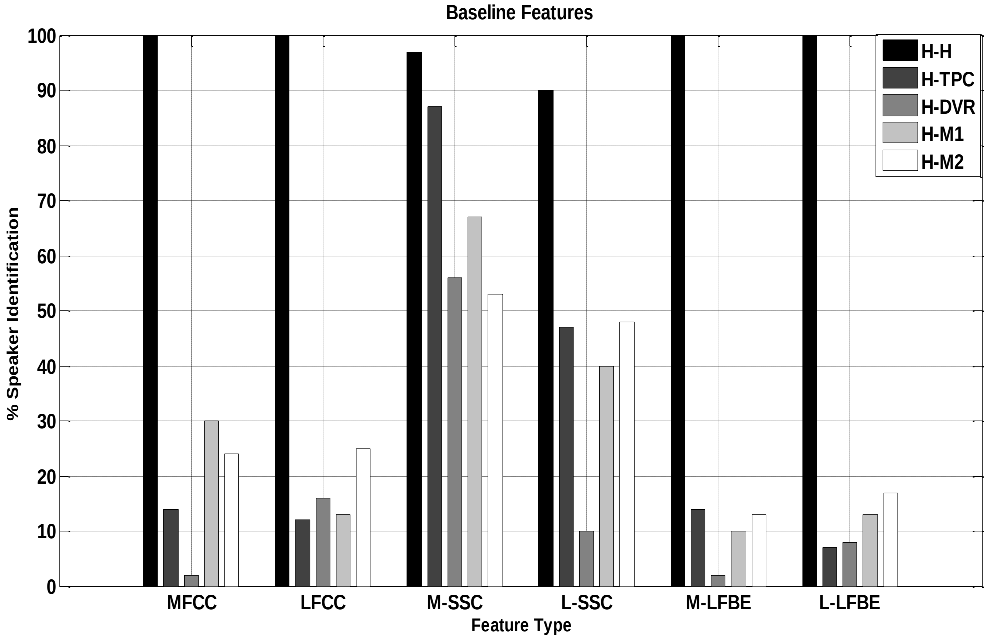

Fig.2. Performance of Speaker Identification with baseline features

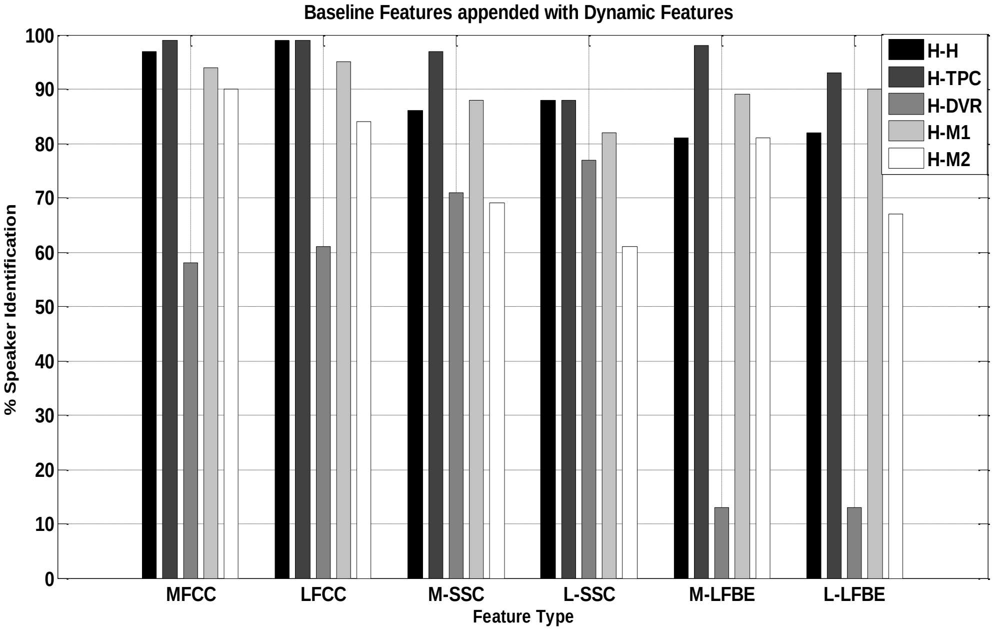

Fig.3. Performance of Speaker Identification with dynamic features

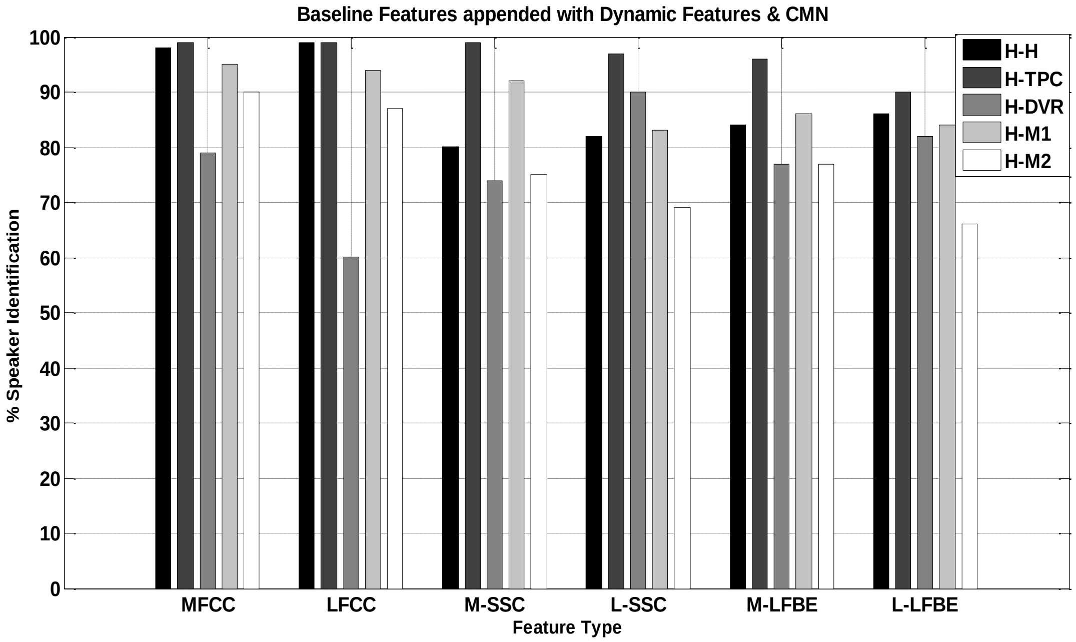

Fig. 4. Performance of Speaker Identification using CMVN

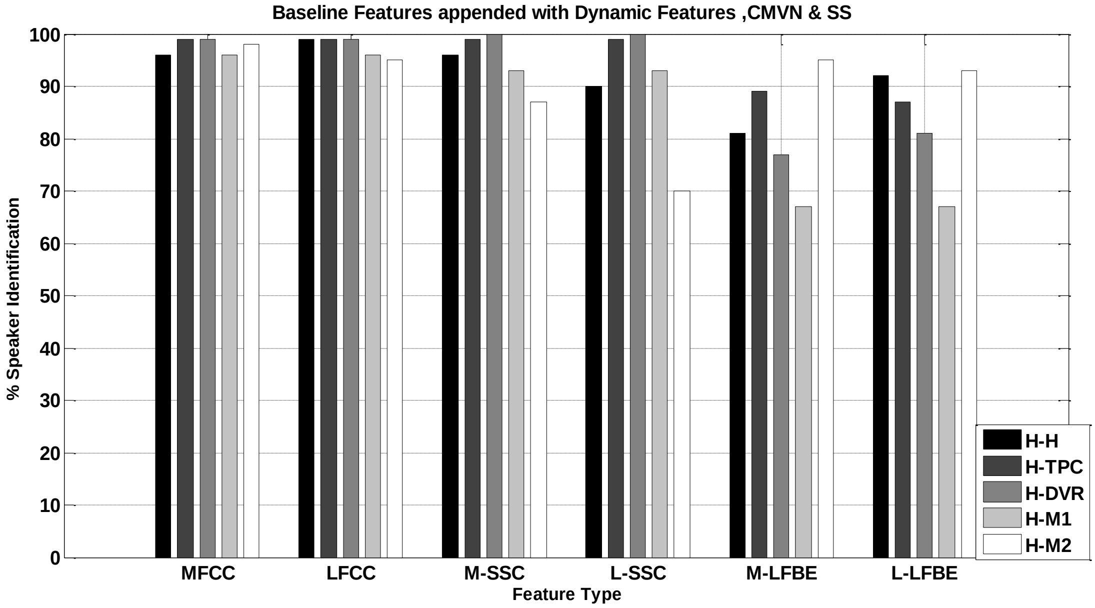

Fig. 5. Performance with CMVN and SS features

6. ACKNOWLEDGMENTS

The authors would like to thank EMST Lab IIT Guwahati for providing speech database.

7. REFERENCES

[1] Joseph P Campbell, Wade Shen, Willam M Campbell,Reva Schwartz, Jean-Francois Bonastre and Driss Matrouf, "Forensic speaker recognition" , *IEEE Signal Processing Magazine,* March 2009 , pp. 95-103.

[2] Tomi Kinnunen, Haizhou Li, "An overview of text independent speaker recognition, from features to supervectors" , *Speech Communication,* July 2009.

[3] Douglas A Raynolds, "Automatic speaker recognition using Gaussian Mixture Model" , *The LINCON Laboratory Journal*, vol.8, No.2, 1995, pp.173-192.

[4] D.A. Reynolds, T.F. Quateri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing* , vol. 10,2000, p. 19-41.

[5] Taufiq Hasan and John H.L. Hansen, "A study of universal background model training in speaker verification", *IEEE Trans. Audio Speech Lang. Process.* vol. 19, No. 7, Sep 2011.

[6] B. Yegnanarayana, and S.P. Kishore, "AANN An alternative to GMM for pattern recognition", *Neural Networks* , vol. 15, 2002, p. 459-69.

[7] V. Wan, and S. Renals, "Evaluation of kernel methods for speaker verification and identification", *Proceeding IEEE International Conference on Acoustic, Speech, Signal Processing.* , vol. 1, 2002, pp.669 –672.

[8] Chang Huai You , Kong Aik Lee and Haizhou Li, "GMM-SVM Kernel with a Bhattacharyya based distance for speaker recognition" , *IEEE Transaction on Audio, Speech and Language Processing*,vol.18,no.6, August 2010, pp.1300-1312.

[9] Marc Ferras, Cheung-Chi Leung, Claude Barras and Jean-Luc Gauvain, "Comparison of speaker adaption methods as feature extraction for SVM-based speaker recognition", *IEEE Transaction on Audio, Speech and Language Processing*,vol.19,no.7, September 2011,pp.1890-1899.

[10] Seyed Omid Sadjadi and John H.L. Hansen, "Robust front end processing ", IEEE ICASSP 2013, pp.7214-7218.

[11] James G Lyons, James G. O'Connel and Kuldip K Paliwal, "Using long-term information to improve robustness in Speaker Identification", *IEEE* 2010.

[12] Xiaojia Zhao and DeLiang Wang, " Analyzing noise robustness of MFCC and GFCC features in speaker identification", IEEE, ICASSP 2013, pp.7204-7208.

[13] Vikramjit Mitra , Mitchel McLaren,Horacio Franco, Martin Graciarena, Nicolas Scheffer, "Modulation features for noise

[14] robust speaker identification", INTERSPEECH 2013, pp. 3707-3713.

[15] Steven V Devis and Paul Mermelstein, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences", *IEEE Transaction on Audio, Speech and Language Processing*,vol.4, ISSP-28,no.4, August 1980 , pp.357-366.

[16] K. K. Paliwal, " Spectral Centroid Features for speech recognition" , Proc. ICASSP, vol. 2, Seattle, 1998, pp.617–620.

[17] Jinggong Chen, Yiteng Huang, Qi Li and Kuldip Paliwal, " Recognition of noisy speech using dynamic spectral subband centroids", IEEE Signal Processing Letters, vol.11, no.2. February 2004,pp. 258-261.

[18] Tomi Kinnunen, Evgeny Karpov and Pasi Franti, " Real time speaker identification and verification", IEEE Transaction on speech and audio processing, vol. 14, no.1, January 2006, pp.277-288.

[19] Electro Medical and Speech Technology Laboratory, Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati. http://www.iitg.ernet.in/ece/emstlab/

[20] Pujol P, Macho D., Nadeu C:On real time mean and variance normalization of speech recognition features,IEEE,ICASSP, 2006

[21] Saeed V. Vaseghi : Advanced Digital Signal Processing and Noise Reduction, Second Edition, John Wiley & Sons Ltd,2000.