Speaker Identification in Mismatch Condition using Warped Filter Bank Features

Mahesh S Chavan
Electronics Engineering Department,
KIT's College of Engineering,
Kolhapur, Maharashtra, India

Sharada V Chougule
Finolex Academy of Management and
Technology, Ratnagiri,
Maharashtra, India

Abstract— Human speech carries variety of information along with actual message. One of these information is identity of the person from his speech. Features carrying the speaker related characteristics are derived from a set of psycho-acoustically motivated filter bank. These filter banks are designed to mimic the human auditory system. The cepstral features derived from these filter bank are used to model the speech or speaker depending upon the task or application. Over the years, MFCCs (mel frequency cepstral coefficients), derived from mel-scale warped filter bank are being used as a physiological feature representing human vocal tract characteristics. Degradation of system performance in any type of mismatch is main drawback of MFCCs. Mismatch in training and testing adversely affects these features and in turn the performance of the system. Therefore extracting the accurate features present in the speech signal related to the speaker, is still a challenging task in mismatched condition. In view of this, alternate frequency warping techniques such as Bark and ERB rate scale can have comparable or better performance to that of mel-scale warping. In this paper the performance short time cepstral features generated using filter banks with Bark and ERB rate warping is investigated in relation to robustness for speaker identification in mismatch condition. For this purpose, two types of sensor mismatched databases are used. Performance of closed set speaker identification system is analyzed under text-dependent and text-independent cases. Front end signal processing is performed with spectral subtraction to reduce the effect of any additive noise. Also normalization of feature vectors is carried out over each frame, to compensate for channel mismatch. Results shows that, Bark and ERB warped filter bank features gives comparable performance to that of mel-scale in text dependant case for both matched and mismatched condition. Whereas bark scale cepstral features having superior performance in mismatch condition.

Keywords— Text-dependant/Text-independent speaker identification, MFCCS, Spectral Subtraction, Mel, Bark , ERB scale

INTRODUCTION

Human listeners uses several sources of information to identify a person from his voice under varying conditions and contexts. The ability of machine to do so is limited by variety of mismatch condition. As human speech is becoming a most popular form of person identity in applications including security, information retrieval or personalization, accuracy of such systems is a crucial issue. Speaker verification and identification are the two types of end tasks depending upon the application. If the end decision is in terms of accepting or rejecting the identity claim of a person, the task is called speaker verification. And if the end decision is to identify a person of best match amongst a set of known speech database, the result is in terms of speaker identification. In case of closed-set speaker identification, the unknown speaker is from the set of N known speaker's database, whereas open-set speaker identification refers to an input speaker which is outside the N known speaker database. The four basic stages of a speaker recognition system are: i) Feature extraction ii) Pattern formation in terms of speaker model iii) Pattern Matching and iv) Decision making. Each of these stages are equally important for a robust system. (*Robustness* here is in relation to mismatched conditions between training and testing of the system for speaker identification task).

In this paper, performance of text independent speaker identification system using warped filter banks features, is analyzed under mismatch condition. Three different warped scales Mel, Bark and ERB respectively are used to derive the cepstral features related to the speaker. For this purpose, speech data with mismatch in sensors (microphones) is used to test the robustness for text-dependent and text-independent cases respectively.

The paper is organized as follows: Section II carries the discussion of feature extraction and recent research in relation to robust features. In Section III, description of the formation of filter banks with mel, bark and ERB scale warping is given. Results of the experiments are discussed in Section IV. Conclusion based on analysis of results is given in Section V.

I. MOTIVATION

Feature extraction is said to be the heart of state-of- art speaker recognition system. These features vary from low level acoustic features to high level lexical syntactic and prosodic features. Low level features derived from sub-band processing or cepstrum are easy to extract and model. Also small amount of data is sufficient for training and testing [1]. But the disadvantage of these features is that, it is easily affected by noise and any type of mismatch. Various situation in real condition such as changes in communication channel, environment and acoustic mismatch, mimicry by human, any external and internal noise can degrade the low level features easily. Various studies are carried out to remove or attenuate noise from speech signal using variety of wavelets like

SYMLET, Harr and Daubechies [2],[3],[4]. Robustness of the system therefore depends on robustness of the features derived input speech. Cepstral features are low level features obtained from short-time spectrum of speech signal weighted (warped) by a psycho-acoustically motivated filter bank like mel-scale filter bank [5]. Mel scale warping is initially used for speech recognition application and same is followed for speaker recognition [6]. But the information required from the features for both task is conflicting. Efforts to make conventional MFCCs robust for continuous speech speaker recognition is done in [7] to reduce the effect of additive and convolutive noise using spectral subtraction and by normalizing the cepstral features derived from mel warped filter bank. Recently use of various warped filter bank features is done to investigate the robustness of speaker identification system [8]. In view of this, along with the use of conventional mel warping, two other warped scale filter bank features are analyzed for speaker identification task under mismatch condition.

II. WARPED FILTER BANK

The speech signal consists of a large amount of raw data (such as pause between utterances or undesired distortions intercepting the speech) which is actually does not carry any useful information. Feature extraction (with preprocessing of speech signal), is the first stage in speaker recognition system. The main purpose of feature extraction is to eliminate the raw speech data and extract the features which convey some characteristics of the speaker. It is a compact and more suitable representation of raw speech data. In contradictory to speech recognition task, the extracted features should carry speaker specific information in the form of acoustic vectors. For better accuracy of the system (in terms of identification or verification), the extracted features should be robust against undesired distortions such as noise or any type of mismatch between data used for training and testing the system.

Psycho-acoustic studies proved that the basilar membrane, located at the front end of the human auditory system, can be modeled as a bank of overlapping band pass filters, each tuned to a specific frequency (the characteristic frequency) and with a bandwidth that increase roughly logarithmically with increasing characteristic frequency. The bandwidth of these filters are known as critical bands of hearing and are similar in the nature to the physiologically-based filters. Nerves at one point of cochlea (inner ear), responds maximum at one particular frequency and less for nearby frequencies. Therefore shape of band pass filters is generally triangular.

Given the roughly logarithmically increasing width of the critical band filters, about 20 to 24 critical band filters can cover the maximum frequency range of 15000 Hz for human perception[5]. A means of mapping linear frequency to this perceptual representation is through warping using either mel-scale, bark scale or ERB (equal rectangular bandwidth) scale. All three scale are based on human perception mechanism discussed above.

MFCCs (Mel Frequency Cepstral Coefficients) is the most popularly used feature extraction technique in state-of-art speaker recognition systems [6]. MFCCs can be considered as filter bank processing adapted to speech specificities.

The cepstral coefficients representing a feature vector from a set of warped filter bank are derived through following steps:

i) Segmentation of speech signal in frames of 20 to 30 msec (this frame duration is generally selected to extract the vocal tract features of a person).

ii) Window analysis of framed speech, with frame overlap of 10 msec.

iii) Spectrum representation and analysis using FFT (Magnitude only).

iv) Formation of filter bank according to auditory scales and finding energy of each filter on mel-scale warping. We use log because our ear works in decibels.

v) Use DCT for compressing the information and decorrelating the source and filter information (related to source-filter mode of human speech production).

Figure 1 illustrate the steps to obtain warped filter bank cepstral features as discussed above.

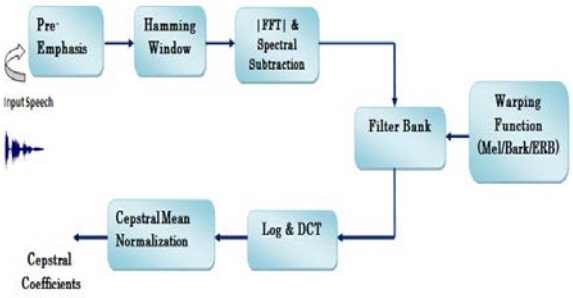


Fig.1 Frequency Warped Filter Bank Feature Extraction

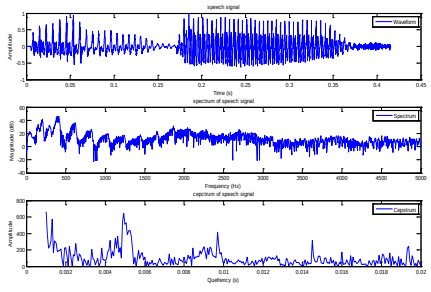


Fig.2 Spectrum and Cepstrum of Speech Signal

III. FREQUENCY SCALES

Frequency scales describe how the physical frequency of an incoming signal is related to the representation of that frequency by the human auditory system. Different psycho-acoustical techniques provides somewhat different estimates of the bandwidth of band pass filters. Mel-scale [6], Bark-scale, ERB (Equal Rectangular Bandwidth)-scale[9],[10] are some of the widely used frequency scales based on frequency domain masking principle.

To analyze the performance of segmental features using perceptually motivated warping, filter banks are designed

using three types on non-linear frequency scales as given below.

A. Mel Scale

Mel-scale frequency is related to linear frequency by empirical equation in (1), which is linear upto 1000 Hz and logarithmic above 1000 Hz. The MEL scale that was proposed by Stevens et al. [11] describes how a listener judges the distance between pitches. The reference point is obtained by defining a 1000 Hz tone 40 dB above the listener's threshold to be 1000 mels.

$$fmel = 2595 * \log 10(1 + \frac{f}{700}) \quad (1)$$

B. Bark Scale

Another means of mapping linear frequency to the perceptual representation is through bark scale. In this mapping one bark covers one critical band with the functional relationship of the frequency *f* to the bark *z* given by [5] as:

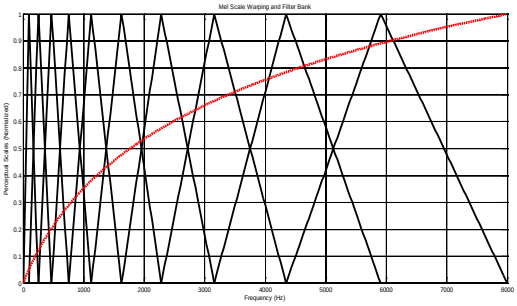
$$z = 13 * \tan^{-1}(0.76 * f) + 3.5 * \tan^{-1}(\frac{f}{7500}) \quad (2)$$

C. ERB Rate Scale

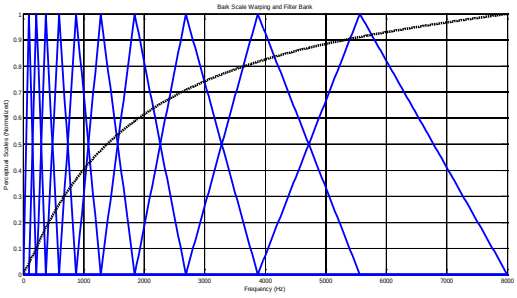
The ERB scale is a measure that gives an approximation to the bandwidth of filters in human hearing using rectangular band pass filters; several different approximations of the ERB scale exist. The following is one of such approximations relating the ERB and the frequency *f*:

$$ERB(f) = 11.17 * \log(1 + \frac{46.065 * f}{f + 14678.49}) \quad (3)$$

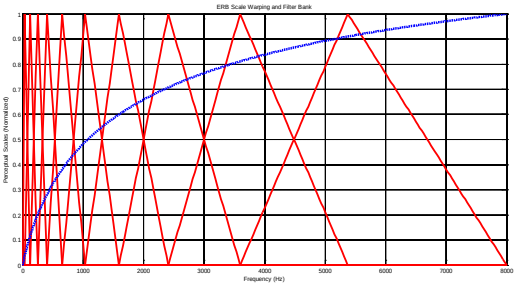
The filter banks designed using above scales for the sampling rate of 16000 Hz are shown below:



(a) Mel-Scale



(b) Bark Scale



(c) ERB Rate Scale

Fig.4 Warped Filter Banks for various Frequency Scales

IV. DATABASE

For performance evaluation, two different database with acoustic mismatch are used.

The first database is of multi-speaker, continuous (Hindi) speech database is used, which is generated by TIFR, Mumbai (India) [12] and made available by Department of Information Technology, Government of India. A set of 100 speakers (TIFR India Hindi Database each speaking 10 different Hindi sentences (of 6 to 8 millisecond) out of which two sentences are same for all speakers. Initially the performance of the system is evaluated for matched conditions (speech data used for training and testing the system is recorded with same close-talking directional microphone).

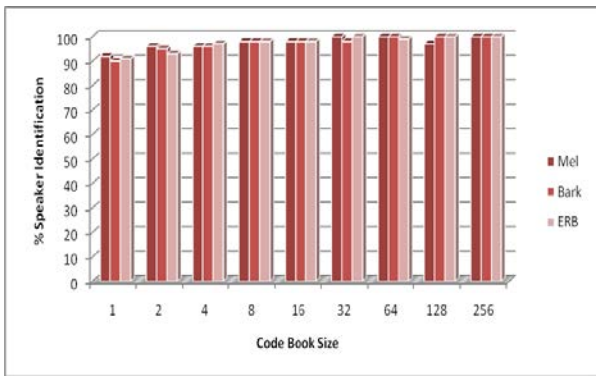
The second database for used is developed by IIT Guwahati. The database consists of four phases, out of which we have used Phase-I, with sensor mismatch. 100 speaker database (81 male and 19 female) whose speech is recorded with three different microphones namely Headset microphone, Table PC build-in microphone, and Digital voice recorder each is recorded with different sampling rate.

V. PERFORMANCE EVALUATION IN MATCHED AND MISMATCHED CONDITION

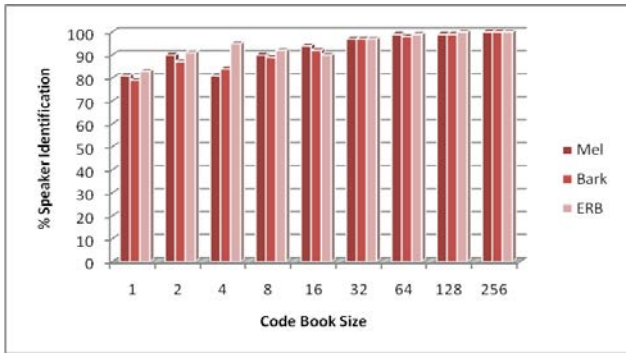
Various undesired distortions like background noise, channel noise damages the quality of speech data when recorded with distant microphone. Experimental evaluations for noisy and multi-speaker environment for speaker recognition is carried out in [13], using LP (linear prediction) based combined spectral and temporal processing approach. A detailed study and overview of feature extraction methods in real world conditions is discussed in [1]. Studied in [14], proposes various alternative mel frequency warped feature representations in the presence of interfering noise.

A closed-set Speaker Identification system is developed in which cepstral features are derived as discussed in section II. To form the model of each speaker, Linde-Buzzo-Gray (LBG) algorithm [15] is developed with codebook size of 128.Thus 39x128 dimensional codebook is formed for each speaker.

The results of percentage identification rate are given for Text-Dependent and Text-Independent case respectively.



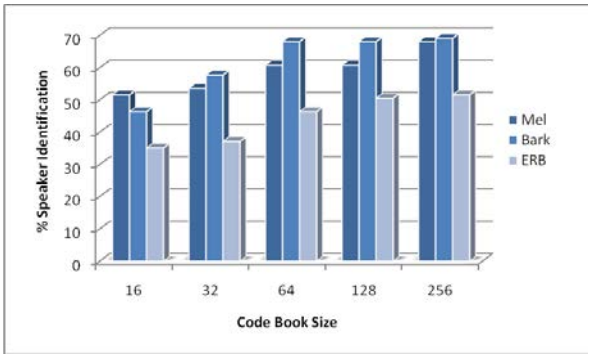
(a) Text Dependent Speaker Identification in Matched Conditions



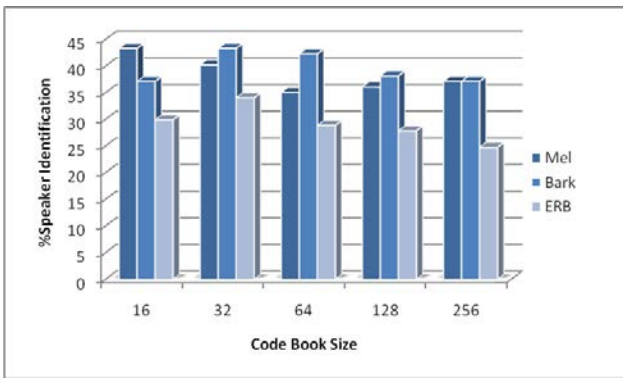
(b) Text Independent Speaker Identification in Matched Condition

Fig.5 Performance of Speaker Identification System for various warping scale for (a) Text-Dependent Case (b) Text Independent Case under *Matched Conditions*

From above graphs, it is observed that, there is negligibly small differenxe in percentage identification rate for text-dependent and text-independent speaker identification under matched conditions.



(a) Text Dependent Speaker Identification in Mismatched Condition



(b) Text Independent Speaker Identification in Mismatched Condition

Fig.6 Performance of Speaker Identification System for various warping scale for (a) Text-Dependent Case (b) Text Independent Case under Mismatched Conditions

From the plots in figure (6), it is observed that, under mismatched conditions, the percentage identification rate decreases for both text-dependent as well as text-independent case. The drop in identification rate is more in text-independent case, than in text-dependent case.

For text-dependent case, under mismatched conditions the reason for decrease in error rate is due to changes in transducer device for training and testing (training with closed talking directional microphone and testing with desk mounted omni-direstional microphone). As the phonetic contents are same for training as well as testing, the only reason for decrease in identification rate is due to channel mismatch and may due to some noise intercepted during testing phase (due to desk mounted microphone).

In text-independent case, the identification rate drop off further. Here the additional reason may be due to changes in phonetic contents during training and testing sets.

VI . COMPENSATION AGAINST MISMATCH

The Cepstral Mean Normalization is often used during the feature extraction phase of speaker recognition/verification systems to compensate for convolutional channel distortion of voice signals. The convolutional channel distortion can be caused by different microphones used between the testing and training phases or different transmission channels used during testing and training. While the Cepstral Mean Normalization is relatively effective in removing convolutional distortion, it is not able to compensate for additive channel distortion.Therefore, it is not capable of removing an additive channel distortion such as white Gaussian noise, which occurs commonly in transmission channels [16].

Cepstral mean normalization (CMN) is an alternate way to high-pass filter cepstral coefficients. In cepstral mean normalization the mean of the cepstral vectors is subtracted from the cepstral coefficients of that utterance on a sentence-by-sentence basis as:

$$y(n) = c(n) - \frac{1}{N}\sum_{n=1}^{N} c(n) \quad (4)$$

The magnitude or power estimate obtained with STFT is susceptible to various types of additive noise (such as background noise). To compensate for additive noise and to restore the magnitude or power spectrum of speech signal, spectral subtraction is used. Magnitude of the spectrum over short duration (equal to frame length) is obtained by eliminating phase information. Here spectrum of noise is subtracted from noisy speech spectrum, therefore it is known as spectral subtraction. For this, noise spectrum is estimated and updated over the periods when signal is absent and only noise is present [17]. Thus speech signal is enhanced by eliminating noise.

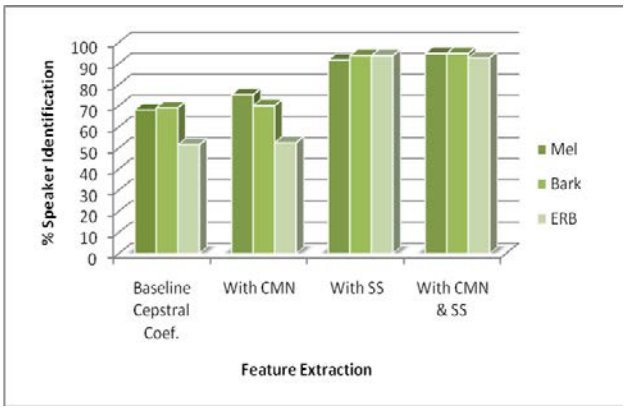
The noisy signal model in the time domain is given by:

$$y(n) = x(n) + n(m) \quad (5)$$

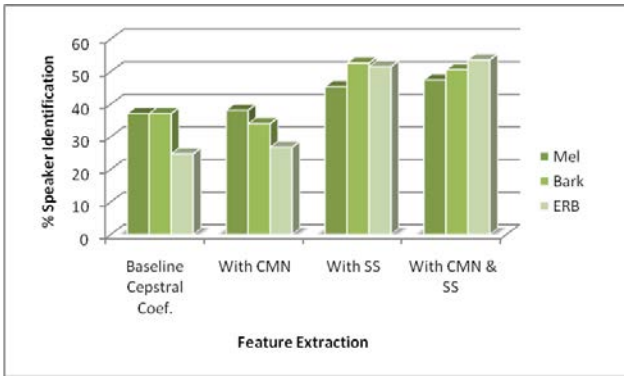
The magnitude spectrum subtraction is defined as:

$$|X(f)| = |Y(f)| - \overline{|N(f)|} \quad (6)$$

where $\overline{N(f)}$ is the time-averaged magnitude spectrum of the noise.



(a) Text Dependent Speaker Identification with compensation against Mismatched Condition

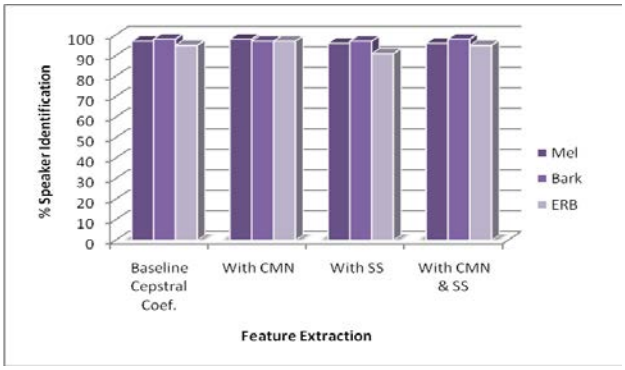


(b) Text Independent Speaker Identification with compensation against Mismatched Condition

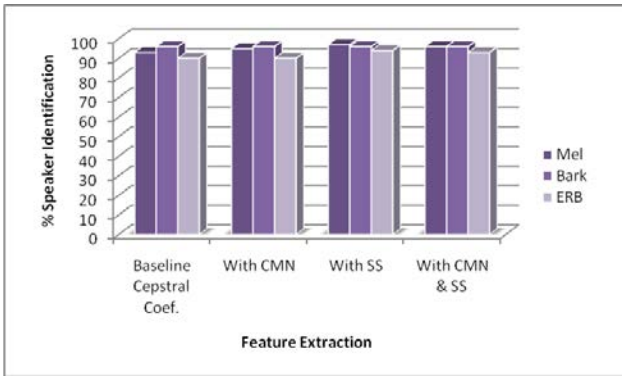
Fig.7 Performance of Speaker Identification System for various warping scale, under Mismatched Conditions using compensation: (a) Text-Dependent Case (b) Text Independent Case

From experimental observations as shown in figure (7), the percentage identification rate is incresed cosiderably by inclusion of spectral subtraction before passing the segmented speech through warped filter banks. It is observed that, normalization of cepstral features have shown negligible improvement in correct identification.

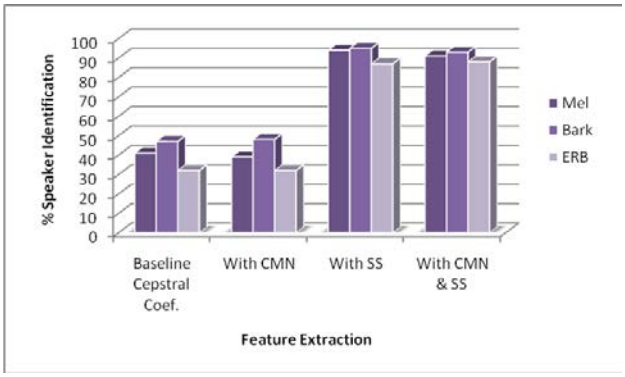
Using the second database of sensor mismatch, with speech recorded using three different microphones, the results of percentage correct speaker identification are shown in figure (7). Here the system is trained and tested for text independent case only.



(a) Training-Headset, Testing-Headset



(b) Training-Headset, Testing-Table PC



(c) Training-Headset, Testing-Digital Voice Recorder

Fig.8 Text Independent Speaker Identification with compensation against mismatch

As observed from plots in figure (8), under matched condition (same sensor for train and test), all three warped filter bank cepstral features gives almost equal performance. In case of mismatched conditions (set b and set c in figure 7), mel scale and bark scale shows almost comparable results, with bark scale warped cepstral features are slightly superior than mel scale filter bank features.

VI. CONCLUSION

In this paper, the performance features based on three frequency warped filter banks is studied for text-dependent and text-independent cases under matched and mismatched conditions.

For *matched conditions* it was observed that, performance (percentage identification rate) is almost identical for both text-dependent and text-independent case. Also the all the three warped filter bank features gives analogous performance for codebook size above 32.

Under *mismatched conditions*, identification rate for text-independent case in much less than for text-independent case. Also, it is observed that, bark warped filter bank features gives somewhat improved performance as compared to mel and ERB warped filter banks, for both text-dependent and text-independent case.

Using spectral subtraction for compensation against any additive noise and normalizing the warped filter-bank features for compensation of mismatch between sensors, the performance of speaker identification system is improved to a better extent. It is observed that, normalizing the filter bank features shows only a little improvement in percentage identification. It indicates that extent of additive noise in acoustic mismatch is more than channel noise. Subtracting the magnitude spectrum estimates the noise interference due to sensor mismatch and eliminated that noise by subtracting noise spectrum from spectrum of noisy speech.

In all, it is concluded that, Bark and ERB warped filter bank features gives comparable performance to that of mel-scale in acoustic mismatch with bark scale cepstral features having superior performance in sensor mismatch.

ACKNOWLEDGMENT

The author would like to thank Dr. Samudravijaya of TIFR Mumbai and EMST Laboratory, IIT Guvahati for providing speech database.

REFERENCES

[1] Tomi Kinnunen, Haizhou Li, An overview of text independent speaker recognition from features to supervectors, Speech *Communication* volume 52, pp 12-40, 2010.

[2] Mahesh S Chavan, Manjusha N Chavan, M S Gaikwad, Studies on Implementation of Wavelet for Denoising Speech Signal, International Journal of Computer Applications Vol. 3,No.2, pp.1-7,June 2010,

[3] Mahesh S. Chavan, Nikos Mastorakis, Manjusha N. Chavan, M.S. Gaikwad, Implementation of SYMLET Wavelets to Removal of Gaussian Additive Noise from Speech Signal, Proceedings of the 10th International conference on Recent Researches in Communications, Automation, Signal Processing, Nanotechnology, Astronomy and Nuclear Physics. Pp 37-41.

[4] Mahesh S. Chavan, Nikos Mastorakis, Studies on Implementation of Harr and daubechies Wavelet for Denoising of Speech Signal, Procedings of International Journal Of Circuits, Systems And Signal Processing Issue 3, Volume 4, pp 83-96,2010 .

[5] Thomas F Quateier,"Discrete Time Processing of Speech Signals-Principles and Practice", Pearson Eduaction,1997.

[6] S. Davis and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. 28, no. 4, pp. 357–366, August 1980.

[7] Chougule Sharada V., Chavan Mahesh S., Channel Robust MFCCs for Continuous Speech Speaker Recognition, Advances in Signal Processing and Intelligent Recognition Systems. AISC, vol. 264, pp. 557–568. Springer, Heidelberg (2014).

[8] Chougule Sharada V., Chavan Mahesh S., Comparison of Frequency Warped Filter Banks in Relation to Robust Features for Speaker Identification, Procedding of 13th International Conference of Signal Processing, SIP'14, Istanbul,Turke, pp 157-162 , December, 2014.

[9] B. C. J. Moore and B. R. Glasberg, A revision of Zwicker's loudness model, *Acustica – Acta Acustica*, Vol. 82, pp. 335–345, 1996.

[10] D O'Shaughnessy, Speech Communication: Human and Machine, Addison-Wesley, 1987.

[11] J. Volkmann, S. S. Stevens, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch (A)," J. Acoust. Soc. Am., vol. 8, no. 3, pp. 208–208, 1987

[12] Samudravijaya K, P.V.S.Rao,S.S.Agrawal, "Hindi Speech Database", Proceedings of International Conference on Spoken Language Processing, 2000, China.

[13] P Krishnamoorthy and S R Mahadeva Prasanna, Application of combined spectral and temporal processing methods for speaker recognition under noisy, reverberant or multi-speaker environment, *Sadhana* Indian Academy of Sciences Vol. 34, Part 5, October 2009, pp. 729–754.

[14] Kenny, J. Cernocky, D. O'Shaughnessy, Frequency Warping and Robust Speaker Verification: A Comparison of Alternative Mel-Scale Representations", *Proc. Interspeech 2013*, pp. 3122--3126, Lyon, France, August 2013

[15] Y. Linde, A. Buzo, and R. M. Gray,"An algorithm for vector quantizer design", IEEE Trans. Commun., vol. 28, no. 1, pp. 84–95, 1980.

[16] X. Menendez-Pidal, R. Chan, D. Wu, and M. Tanaka, "Compensation of channel and noise distortions combining normalization and speech enhancement techniques", Speech Communication vol. 34, pp. 115–126, 2001.

[17] Saeed V. Vaseghi ,"Advanced Digital Signal Processing and Noise Reduction, Sec-ond Edition, John Wiley & Sons Ltd,2000

Mahesh S Chavan has received B.E.(Electronics Engineering) in 1991 from shivaji University and Master of Engineering in Electrical Control Systems from Shivaji University, Kolhapur India in 1996. He has received Ph D in Electronics from Kurukshetra University, Kurukshetra India. He has 24 years of teaching experience. Presently he is working as a Professor in Department of Electronics Engineering, K I T's College of Engineering, Kolhapur. His area of research interest includes study of Signal Behaviour, Nonlinear Control Systems.

Sharada V Chougule has received B.E (Electronics Engineering) in 1992 from Shivaji University Kolhapur and M E (Electronics Engineering) from Shivaji University, Kolhapur in 2010. Presently she is perusing Ph D from Department of Technology Shivaji University, Kolhapur. She has 18 years of teaching experience. Her areas of interest include Digital Signal Processing, Speech processing, Speech and Speaker recognition and Biometric recognition.