SPEAKER IDENTIFICATION IN NOISY ENVIRONMENT

Naresh P. Jawarkar
B. N. College of Engineering, Pusad (MS), India

Abstract

An accurate speaker identification is difficult due to a number of factors. One of the most prominent factors is environmental noise. In this paper, the effect of additive noise on the performance of the closed set text-independent speaker identification system is addressed. Performance is analyzed with Mel frequency cepstral coefficients (MFCC) and Gammatone frequency cepstral coefficients (GFCC). The effect of two nonlinear compression functions, namely log and cubic root used in the feature extraction process of GFCC on the system performance is analysed. The Gaussian mixture model approach is for speaker modelling. Two databases, namely Marathi and Hindi databases were used for the experimentation. It has been observed that MFCC based system performs better than GFCC based system in clean and low noisy environment. Whereas GFCC based system outperforms MFCC based system in highly noisy environment.

Index Terms: Speaker identification, MFCC, GFCC, Noisy environment.

I. INTRODUCTION

The area of speaker recognition is concerned with extracting the identity of the person speaking. One of the advantages of using speech to determine an individual's identity is that speech is the most natural means of interacting with each other. Speaker recognition task can be classified into two parts, viz. speaker identification (SI) and speaker verification (SV). The former refers to determining who is talking from a set of known voices or speakers. The latter refers to determining whether a given voice sample is produced by a claimed speaker. Speaker identification task can further be classified into open- and closed-set tasks. If the target speaker is assumed to be one of the registered speakers, the recognition task is a closed set problem. If there is a possibility that the target speaker is none of the registered speakers the task is called open-set problem. Speaker identification can also be classified into text-dependent and text-independent tasks. In former the utterance presented to the recognizer is known before hand. In the later case, no assumptions about the text being spoken, is made. At the fundamental level, automatic speaker recognition is a pattern recognition task which consists of two main blocks namely, the feature extractor and a pattern classifier. The feature extractor does the job of mapping the speech signal into set of feature called feature vector, and in the speaker pattern classifier module the test sample from the target speaker is compared against the speaker database. The comprehensive review of speaker recognition is given in [1]-[4].

Classical speaker models can be divided into template models and stochastic models, also known as non parametric and parametric models, respectively. Vector quantization (VQ) [5] and dynamic time wrapping (DTW) [6] are representative examples of template models for text-independent and text-dependent recognition, respectively. The Gaussian mixture model (GMM) [7], [8] and the hidden Markov model (HMM) [9]-[11] are the most popular stochastic models for text-independent and text-dependent recognition, respectively. According to the training paradigm, the models can also be classified into generative and discriminative models. The generative models such as GMM, VQ and Probabilistic Neural Network (PNN) [12] estimate the feature distribution within each speaker. The discriminative models such as artificial neural networks (ANN) [13]-[14] and support-vector machines (SVM) [15] in contrast,

model the boundary between speakers. The other classifiers include Polynomial classifier [16], [17] and k-nearest neighbor classifier [18], Fuzzy min-max neural network [19]-[20], sparse representation classifier [21].

Speech signal includes many features which can be used for speaker discrimination. Features can be categorized into short-term spectral features, voice source features, spectro-temporal features, prosodic features, high level features, etc. [3]. Low-level features are generally related to physical traits of a speaker's vocal apparatus. Short-term spectral features, as the name suggests, are computed from short frames of about 20-30 milliseconds in duration. They generally describe the short-term spectral envelope which is an acoustic correlate of timbre as well as the resonance properties of the supra-laryngeal vocal tract [4]. The short term features include mel-frequency cepstral coefficients (MFCC) [22], super-mel-frequency cepstral coefficients [23], linear predictive cepstral coefficients (LPCC) [24], Perceptual linear prediction (PLP) coefficients [25], and Line spectrum frequencies (LSF) [26], auditory features [27], etc.

Automatic speaker recognition can achieve a high level of performance in matched training and testing conditions. However, such performance drops significantly in mismatched noisy conditions. It has been observed that the accurate speaker recognition is difficult due to a number of factors. Two most prominent factors are handset/channel mismatch and environmental noise. Much research has been conducted with a focus on reducing the effect of handset/channel mismatch. Linear and nonlinear compensation techniques have been proposed, with applications to feature, model and match-score domains. Examples of the feature compensation methods include well-known filtering techniques such as cepstral mean subtraction or RASTA [28]-[30]. Score-domain compensation aims to remove handset-dependent biases from the likelihood ratio scores. The most prevalent methods include H-norm [31], Z-norm [32], and T-norm [33]. In conventional MFCC computations, *Log* is used as nonlinear compression function. Root compressed spectral approaches represent speech better in noise [34]-[36].

The main objective of the present work is to carry out the comparative study of two feature sets, namely, MFCC and GFCC for the close-set text-independent speaker identification under noisy environment using Gaussian mixture models. GMM is currently one of the principal methods for speaker identification.

Rest of the paper is organized as under. Section-II deals with the feature extraction process. Experimental analysis and results are given in section III. Finally conclusion is discussed in section IV.

II. FEATURE EXTRACTION

The procedure for front end processing, which includes speech recording, pre-processing and feature extraction, is depicted in the block schematic shown in Fig. 1. The speech signal is first passed through anti-aliasing filter with cut-off frequency of 44.1 KHz. The signal is then sampled at the sampling frequency of 22050 Hz and converted into digital signal using analog to digital converter with 16-bit resolution. The non-voiced portion of the signal is removed by the silence removal stage using the energy threshold criterion. Feature extraction process for MFCC and GFCC is discussed below.

A. Mel frequency cepstral coefficients

The voiced speech signal, after silence removal, is pre-emphasized with pre-emphasis factor of 0.97. This is followed by frame blocking with a frame length of 512 samples with 50% overlap with the neighboring frames. Finally each frame is multiplied with Hamming window to reduce the side lobe effects. Magnitude spectrum of each frame is obtained by taking FFT. This spectrum is multiplied by the mel-scale triangular filters and then log energy is computed. The log-energy filter outputs are then cosine transformed to produce the cepstral coefficients. The zeroth cepstral coefficient is discarded as it contains only DC term.

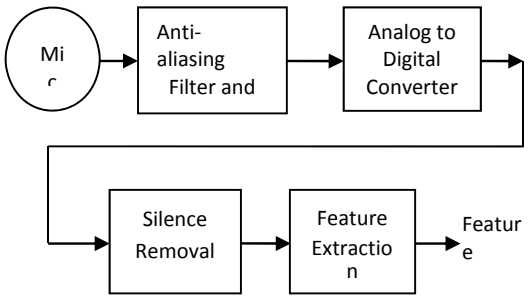

Fig. 1- Speech recording and feature extraction

B. *Gammatone frequency cepstral coefficients*

Humans are found to perform better than machines in speaker recognition tasks when input signals are corrupted by background noise [37]. To tackle this robustness problem, Y. Shao, et al. introduced a novel speaker feature, Gammatone frequency cepstral coefficients, based on an auditory periphery model [27]. Gammatone filters are derived from psychophysical and physiological observations of the auditory periphery and this filterbank is a standard model of cochlear filtering [38]. The impulse response of a Gammatone filter centred at frequency *f* is:

$$g(f_c,t) = \begin{cases} t^{a-1}e^{-2\pi bt}\cos(2\pi f_c t), & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (1)$$

Where f_c is the centre frequency, a is the order of filter, and b is the rectangular bandwidth. Patterson [38] suggested that a 4th order Gammatone filter would be a good model of the auditory filter.

Different steps involved in GFCC computation are:

- The input signal is passed through a Q-channel gammatone filterbank, followed by the equal loudness stage.
- The filtered signals are divided into frames of 23.22 ms length with 50 % overlapping. Energy in each frame of the subband signal is computed.
- A nonlinear compression function(Cubic root or Log) is applied to energy of each frame. In standard GFCC cubic-root function is used.
- Finally, DCT is applied to derive cepstral features.

Thirty filters were used GFCC computation.

C. *Cepstral mean subtraction*

It is desirable to design a feature extractor which normalizes the features before feeding them onto the modelling or matching algorithms, particularly for robust speaker identification. The simplest method of feature normalization is to subtract the mean value of each feature over the entire utterance. This is known as cepstral mean subtraction (CMS) [24], [39].

$$\vec{c}(n) = c(n) - \mu \quad (2)$$

where $C(n)$ is the feature vector of the n^{th} frame and μ is the mean of feature vectors and is given by:

$$\mu = \frac{1}{T}\sum_{n=1}^{T} c(n) \quad (3)$$

where T is number of frames in the analyzed speech signal.

III. EXPERIMENTATION

A. *Database used for experimentation*

Two databases, namely Marathi-database and Hindi-database were employed in this study. The databases consist of speech utterances containing words, digits and sentences.

Marathi Database: In this case, speech utterances were recorded in Marathi- a language spoken over by 8 crores of population in the state of Maharashtra and the neighboring states in India. Speeches were recorded for eighty speakers in the age group of 8 to 72 years.

Hindi Database: This database consists of speech utterances consisting of words, digits and sentences in Hindi language. The speech utterances were stored using a digital recorder for forty two speakers in the age group of 18 to 23 years in the Jharkhand state of India.

B. Experimental Results

The speaker identification mainly consists of speaker enrolment and pattern recognition. Clean speech utterance of 1 minute duration was used for training for each speaker. Each speaker was modelled using GMM with 16 mixtures. Expectation-Maximization algorithm was employed for GMM-training. Identification performance for each classifier was carried out for 1, 3, 5, and 10 second test utterance lengths of clean and noisy speech. Noisy speech test utterances were obtained by adding six different noise signals namely, White Gaussian noise (WGN), Small factory ambiance (SFA), Lathe machine noise (LMN), Miller machine noise(MMN), Grinder machine noise (GMN) and Shaper machine noise(SMN) to clean speech signal with different values of signal to noise ratio (SNR). The test speech was first processed to evaluate the sequence feature vectors. The sequence of feature vectors was divided into overlapping segments of feature vectors at the 23.2 ms frame rate. Thus, 1-second testing utterance contains 86 feature vectors. Speaker identification accuracy (SIA) is calculated as:

$$SIA = \frac{1}{N}\sum a_i \quad (4)$$

Where a_i is the speaker identification accuracy of the i^{th} speaker and is defined as under:

$$a_i = \frac{N_C}{N_T} \times 100 \quad (5)$$

Where N_C is the number of correctly identified segments and N_T is the total number of test segments for the i^{th} speaker.

The selected values of the parameters used in MFCC processing are shown in Table I.

Table I: Front end processing details for MFCC

Parameter	Value
Number of Cepstral coefficients	18
Number of filters	24
Frequency range for filters	100 Hz - 4000 Hz
Frame size	512 samples (23.22 ms)
Frame rate	256 samples (11.61 ms)

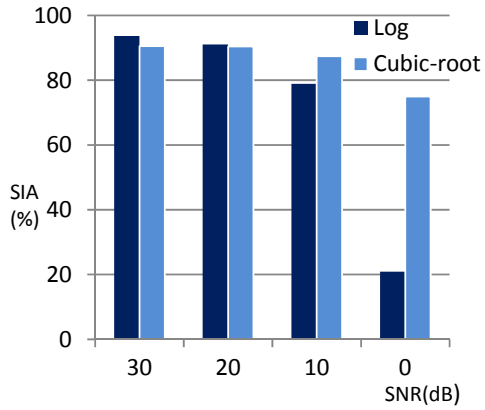

Fig. 2- Performance with Log based and Cubic-root based GFCC (Hindi database)

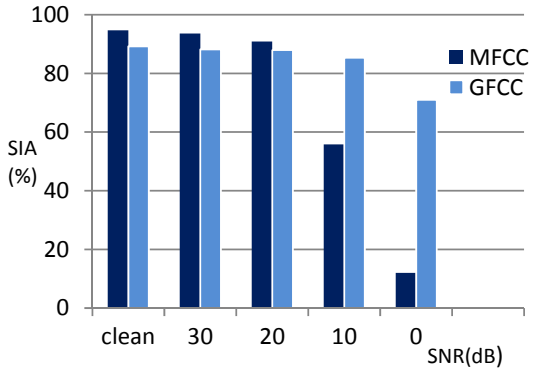

Fig. 3- Performance with MFCC and GFCC features for Hindi database

Fig. 2 shows the performance with GFCC with two nonlinear compression functions for 10 second test speech utterance. It is observed that log based GFCC is effective for the SNR of 20 dB or above. On the other hand, cubic-root based GFCC is effective for the noisy environment with low SNR. Fig. 3 shows the performance of the SI system with MFCC and standard GFCC for different WGN noise levels. It can be seen that the performance with MFCC is better as compared to that with GFCC for SNR more than 20 dB, whereas GFCC outperforms MFCC for highly noisy environment.

Table II: SIA with GFCC for Hindi database with White Gaussian Noise

SNR (dB)	SIA (%)			
	1s	3s	5s	10s
∞	79.56	87.58	89.19	91.17
30	75.07	85.53	88.18	90.58
20	74.95	85.21	87.95	90.42
10	72.12	82.46	85.33	87.39
0	57.19	68.59	70.99	74.96

Rest of the experimentation was carried out with cubic-root based GFCC. Table II to Table VII show the performance of the GFCC based system for six different noises for Hindi database. The performance of the system for different noises with 10 second speech utterance for Hindi and Marathi database are shown in Fig. 4 and Fig. 5, respectively. It can be seen that the accuracy reduces as the signal to noise ratio reduces. Further the accuracy with White Gaussian noise and Shaper Machine noise is better for low SNR as compared to other noise types.

Table III: SIA with GFCC for Hindi database with Small Factory Ambiance noise

SNR (dB)	SIA (%)			
	1s	3s	5s	10s
∞	79.56	87.58	89.19	91.17
30	74.93	85.43	88.15	90.94
20	74.28	84.79	87.56	89.74
10	68.24	80.14	83.87	85.85
0	36.68	45.66	47.97	50.74

Table IV: SIA with GFCC for Hindi database with Lathe Machine Noise

SNR (dB)	SIA (%)			
	1s	3s	5s	10s
∞	79.56	87.58	89.19	91.17
30	79.01	86.91	88.52	90.64
20	74.35	84.82	87.68	89.81
10	68.26	79.58	83.14	85.95
0	31.28	39.13	42.77	45.70

Table V: SIA with GFCC for Hindi database with Miller Machine Noise

SNR (dB)	SIA (%)			
	1s	3s	5s	10s
∞	79.56	87.58	89.19	91.17
30	79.21	86.71	88.33	90.48
20	74.76	84.4	87.30	89.57
10	68.59	80.52	83.46	86.26
0	32.09	41.87	46.06	49.78

Table VI: SIA with GFCC for Hindi database with Grinder Machine Noise

SNR (dB)	SIA (%)			
	1s	3s	5s	10s
∞	79.56	87.58	89.19	91.17
30	79.12	87.03	88.83	90.72
20	74.88	84.46	87.30	89.57
10	65.94	77.30	80.01	82.52
0	35.35	41.71	45.58	49.27

Table VII: SIA with GFCC for Hindi database with Shaper Machine Noise

SNR (dB)	SIA(%)			
	1s	3s	5s	10s
∞	79.56	87.58	89.19	91.17
30	79.03	86.87	88.59	90.67
20	74.58	85.06	87.81	90.12
10	71.80	82.43	85.18	87.47
0	59.68	70.15	73.00	75.14

IV. CONCLUSION

In this paper the issue of speaker identification under noisy environment has been addressed. It is observed that the speaker identification accuracy reduces as the noise level increases. MFCC performs better than GFCC under clean and low noise environment (SNR ≥ 20 dB). The GFCC outperforms the MFCC under highly noisy environment. Effect of nonlinear compression functions: log and cubic-root used in extraction of GFCC feature on the performance has been studied. It has been found that the cubic-root based cepstral features performs better than log based cepstral features under highly noisy conditions. The future work involves designing a combined system, which uses both MFCC-GMM and GFCC-GMM systems and study its performance under clean and noisy test environments.

ACKNOWLEDGEMENTS

Author is grateful to Dr. H. B. Nanvala, Principal, Babasaheb Naik College of Engineering, Pusad for providing necessary facilities towards carrying out this work. Author also thanks Mr. S. P. Pathe for his laboratory support.

REFERENCES

[1] Marcos Fauder-Zanuy and E. Monte-Moreno, "State – of – the – art in speaker recognition," *IEEE A & E Systems Magazine*, pp. 7-12, May 2005

[2] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters,* vol. 18, no. 9, pp.859-872, Sept.1997.

[3] T. Kinnunen and H. Li, "An overview of text_independent speaker recognition:From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp.12-20, Jan 2009.

[4] Q. Jin and T. F. Zheng, "Overview of front-end features for robust speaker recognition," in *Proc. APSIPA,* 2011.

[5] Y.Linde, A.Buzo and M. Gray, "An Algorithm for vector Quantization," *IEEE Trans. on Communication,* Vol. COM-28, no. 1, pp. 84-95, Jan.1980.

[6] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE trans. Acoustic, Speech Signal Process.,* vol. 29, no. 4, pp. 254-272, 1981.

[7] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture models," *IEEE Trans. on Speech & Audio Processing*, vol. 3, no. 1, pp. 72-83, Jan. 1995.

[8] D. A. Reynolds, 'Speaker identification and verification using Gaussian mixture models," *Speech communication*, vol.17, pp. 91-108, 1995.

[9] Che C. W. and Lin Q., "Speaker recognition using HMM with experiments on YOHO database," in Proc. EUROSPEECH, pp 625-628,1995.

[10] N. Z. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. on Speech & Audio Processing*, Vol. 39, no. 3, pp. 563-570, 1991.

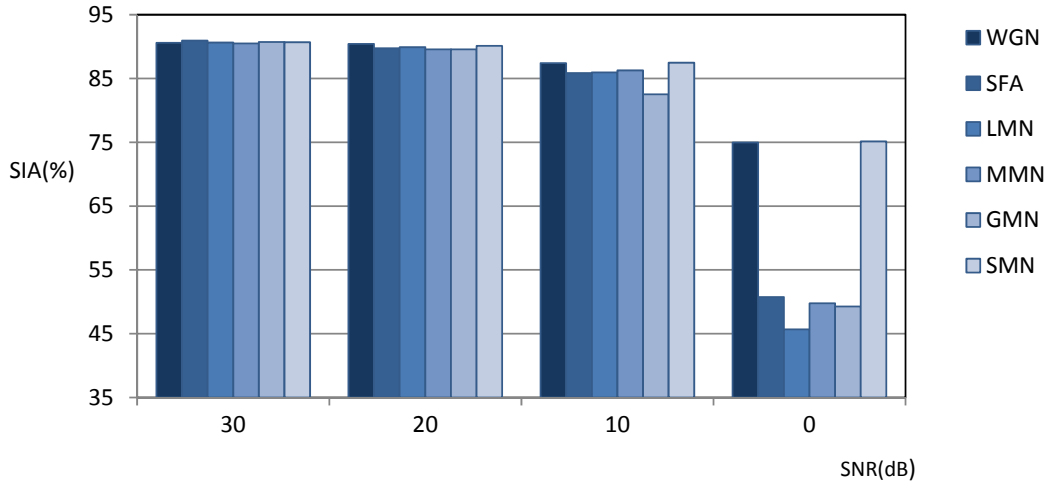

Fig. 4- Performance of speaker identification for different noises with Hindi Database
WGN: white Gaussian Noise, SFA: Small factory Ambiance, LMN: Lathe Machine Noise, MMN: Miller Machine Noise, GMN: Grinder Machine Noise, SMN: Shaper Machine Noise

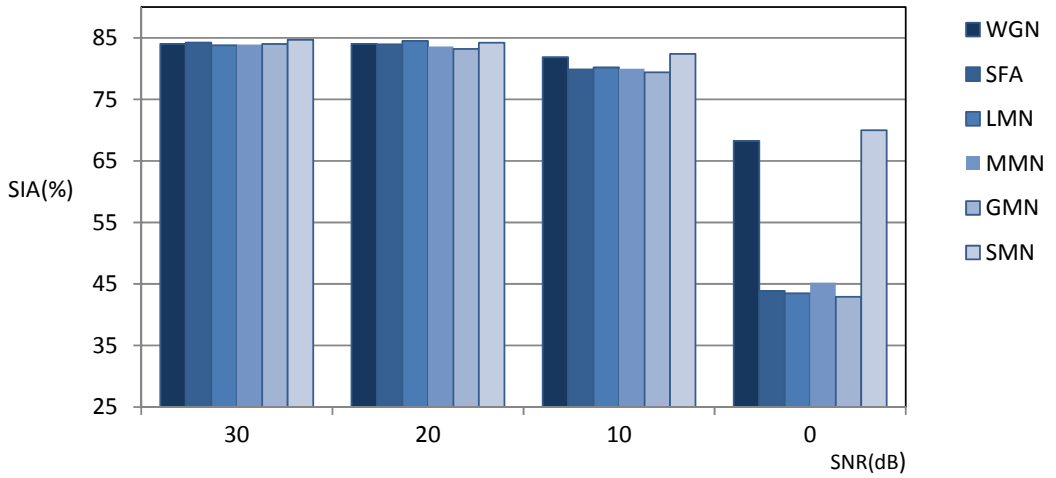

Fig.5- Performance of speaker identification for different noises with Marathi Database

[11]Rabiner L. R., "A tutorial on Hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp 257-286, 1989.

[12]Ganchera T. D., Tasoulisb D. K.,Vrahatisb M. N. and Fakotakisa N. D., "Generalized locally recurrent probabilistic neural networks with application to text-independent speaker verification," *Neuro Computing*, vol. 70, no. 7-9, pp. 1424-1438, March 2007.

[13]Matějka, P. et al, "Analysis of DNN approaches to speaker identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016* pp. 5100-5104, March 2016.

[14]K. R. Farrell, R. J. Mammone and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. on Speech & Audio Processing*, vol. 2, no. 1, part-II , pp. 194-205, Jan. 1994.

[15]Campbell W.,Campbell J.,Raynolds D.,Singer E., "Support vector machines for speaker and language recognition," *Computer Speech Lang.,* vol.20(2-3), pp210-229,2006.

[16]Assaleh K. T. and Campbell W. M. , "Speaker identification using a polynomial classifier," *ISSPA'99*, Brisbane, Australia, pp 115-118.

[17]W.M. Campbell, K.T. Assaleh and C. C. Broun, "Speaker recognition with polynomial classifiers," *IEEE Trans. on*

Speech & Audio Processing, Vol. 10, no. 4, pp. 205-212,May 2002.

[18]A. Higgins, L. Bahler and J. Porter, "Voice identification using nearest neighbor distance measure," in *Proc. IEEE ICASSP*, 1993, pp.H-375-H-378, Apr. 1993.

[19]P. K. Simpson, "Fuzzy min-max neural network-Part 1: Classification," *IEEE trans. Neural Networks*, vol.3,pp.776-786, Sept. 1992.

[20]N. P. Jawarkar, R. S. Holambe and T. K. Basu, "Use of Fuzzy Min-Max Neural Network for Speaker Identification," in *IEEE International Conf. on Recent Trends in Information Technology (ICRTIT),* pp.178-182, June 2011.

[21]Y. H. Chin, et al., "Speaker Identification Using Discriminative Features and Sparse Representation," *IEEE Transactions on Information Forensics and Security* , vol. 12, no. 8, Aug. 2017.

[22]S. Devis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans Acoustics, Speech, Signal Process*, vol. 28,no. 4,pp. 357-366, 1980.

[23]Ma, Z., Yu, H., Tan, Z. H., & Guo, J., "Text-Independent Speaker Identification Using the Histogram Transform Model," *IEEE Access*, vol. 4, pp. 9733-9739, 2017.

[24]B. S. Atal, "Effectiveness of linear prediction of the speech wave for automatic speaker identification and verification," *J. Acoustical Society of America*, vol. 55, no. 6, pp. 1304-1312, June 1974.

[25]H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech," *J. Acoust. Soc. America*,82, pp.1738-1752, 1990.

[26]F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoustic. Soc. America*, vol. 53, 537(A), 1995.

[27]Y. Shao, et al., "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. ICASSP*-2007, IV, Honolulu, Hawaii, U.S.A, pp. 277-280, 2007.

[28]R. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition-A feature-based approach," *IEEE Signal Process. Mag.*, vol. 13, no. 5, pp. 58–71, Sep. 1996.

[29]H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[30]D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 639–643, Oct. 1994.

[31]D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[32]C. Barras and J. L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. ICASSP'03*, Hong Kong, China, pp. 49–52, 2003.

[33]R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Process.*, vol. 10, pp. 42–54, 2000.

[34]P. Alexandre and P. Lockwood, "Root cepstral analysis: a unified view, application to speech processing in car noise environments," *Speech Commun*., vol. 12, no. 3, pp. 277–288, Jul. 1993.

[35]Zhao, X., & Wang, D., " Analyzing noise robustness of MFCC and GFCC features in speaker identification," In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013),* pp. 7204-7208, May 2013.

[36]Jawarkar, N. P., Holambe R. S, and Tapan Kumar Basu, "Effect of Nonlinear Compression Function on the Performance of the Speaker Identification System under Noisy Conditions." in *Proceedings of the 2nd International Conference on Perception and Machine Intelligence*, pp. 137-144. ACM, 2015

[37]Schmidt-Nielsen, A., and Crystal, T. H., "Human vs. machine speaker identification with telephone speech" in *Proc. ICSLP*, Sydney, Australia, 1998.

[38]Patterson, R. D.*, et al. "*Auditory models as preprocessors for speech recognition," in *The auditory processing of speech: From sounds to words,* M.E.H. Schouten, Ed., Berlin, Germany: Mouton de Gruyter, pp. 67-83, 1992.

[39]Furui, S. "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustic, Speech Signal Process.,* vol. 29, no. 4, pp. 254-272P, 1981.