International Journal For Technological Research In Engineering *ISSN (Online): 2347 - 4718*
International Conference on Emerging Technologies in Engineering, Biomedical, Medical and Science (ETEBMS - November 2016)

MGFCC COMBINED ASR SYSTEM FOR ENVIRONMENT FRIENDLY SPEECH ANALYSIS

Er. Vishnu Rajan[1], Prof. Dr. Yashpal Singh[2], Dr. Swati Sharma[3]
[1]Research Scholar, Dept of ECE, Jodhpur National University
[2]SITM, Rewari, [3]Associate Professor, Dept of EE, Jodhpur National University

Abstract: Maximizing the performance of feature extraction techniques has been a speech synthesizer's primary objective from the beginning of research. The MFCC and GFCC feature components combined are suggested to improve the reliability of a speaker recognition system. The MFCC based speaker recognition provides high accuracy and it is a low complex systems; however they are not very robust at the presence of additive noise. The GFCC features in recent studies have shown very good robustness against noise and acoustic change. The main idea is to integrate MFCC & GFCC features to improve the overall ASR system performance in low signal to noise ratio (SNR) conditions. The experiment are conducted on the English Language Speech Database for Speaker Recognition (ELSDR) databases, were the test utterances are mixed with noises at various SNR levels to simulate the channel change. The results provide an empirical comparison of the MFCC-GFCC combined features and the individual counterparts

Index Terms: MFCC features, GFCC feature, combined system using MFCC and GFCC features.

I. INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking by using the speaker-specific information included in speech waves to verify identities being claimed by people accessing systems. By using this technology we can able to make access control for various services by voice. Applicable services include voice dialing, telephone shopping, information and reservation services, voice mail, security control for highly confidential information, database access services, banking over a telephone network and remote access to computers.Most commonly used application of speaker recognition technology is as a forensics tool. For real world application noise robust automatic speech recognition systems are essential .We have to remove additive noise, room reverberation and channel/handset variations from the received noisy speech signal. Improving the noise robustness has been a research task for many years. To reduce the mismatch between training and test conditions [2] speakers can be modeled in multiple noisy environments. Some of the currently using Speech enhancement methods are spectral subtraction[5], noise-robust speaker recognition .Computational auditory scene analysis (CASA) can be used to remove noise .Speaker features such as modulation spectral features and those incorporating phase information have shown robustness against reverberation. Blind DE reverberation algorithms have been used to restore the anechoic signal or the early reflections of reverberant speech. Borgstrom and McCree modeled the effect of reverberation as a channel-wise convolution of short-time spectral envelopes. The National Institute of Standards and Technology (NIST) has conducted a series of speaker recognition evaluations (SRE) since 1996.State-of-the-art systems include joint factor analysis and i-vector based techniques. DEEP neural networks (DNNs) [1] have been adopted in many Automatic Speech Recognition (ASR) systems [3][7]. Large performance improvements have been reported compared to systems that use Gaussian Mixture Models (GMMs). For noisy speech recognition, DNNs have also obtained comparable performance to the best GMM system with various noise reduction, feature enhancement and model-based compensation methods. However, DNNs are still far from reaching humans' expectations and few methods have been developed to further improve DNNs' noise robustness .To a certain extent, DNNs may be capable of learning some noise-dependent feature normalization effects implicitly through multiple layers of non-linear transformations.

II. SYSTEM OVERVIEW AND FRONT-END PROCESSING

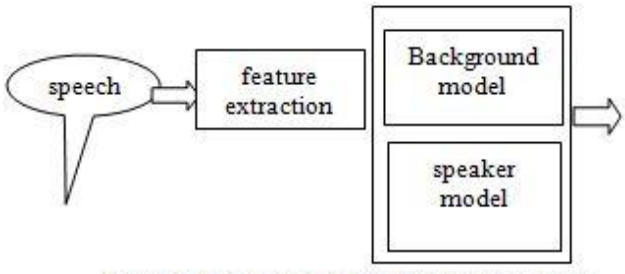

Fig 1: Block diagram of speaker identification system

A. VAD

Signals must be first filtered to rule out the silence part, otherwise the training might be seriously biased. We use a simple energy-based approach to remove the silence part, by simply remove the frames that the average energy is below 0.01 times the average energy of the whole utterance.

B. Feature Extraction

Recent research has shown that the auditory features which shows high performance are GF and GFCC.which promises high robustness to the ASR system than other auditory features such as melfrequency cepstral coefficients (MFCC).

The Gammatone Frequency Cepstral Coefficients (GFCC) are auditory feature based on a set of Gammatone Filter banks.GF vector can be generated by rectification of each frame of the cochleagram using the cubic root operation. GFCC can be derived from GF by applying discrete cosine transform on it [9]. Mel-Frequency Cepstral Coefficient is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency. Linear predictive coding is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. The basic assumption in LPC is that, in a short period, the nth signal is a linear combination of previous p signals

$$x(n) = \sum_{i=1}^{p} ai\ x((n-1)$$

C. Universal background model (ubm)

The task to detect a speaker could be defined as two hypothesis tests. The first test is the one in which the speech signal Z does come from the hypothesized speaker and the second one where it does not come from the hypothesized speaker. The likelihood of the hypothesis Hi given the speech signal can be defined as the probability density function p(Z | Hi). Then we can use a likelihood ratio test given by the two hypotheses to determine the decision. The likelihood function, $(X \mid \lambda)$ selected for calculating the likelihood ratio of the model is very important. For text-independent speaker recognition the most successful one has been the Gaussian mixture models (GMM). A GMM could be thought of as a Gaussian distribution describing a one dimensional random variable X. The variable X is defined as a vector described by the mean and variance. The mixture density for a feature vector, X can be defined as

$$p\ (X \mid \lambda) = \sum_{i=1}^{m} wi\ pi\ (X)z$$

This mixture density is a weighted linear combination of unimodal Gaussian densities, (X)

$$pi\ (X) = \frac{1}{2\pi^{D/2} \mid \sum_{i} 1/2} e^{!/2\ (x-\mu i)^T} \sum_{i}^{-1} (x - \mu i)$$

The UBM is trained using the Expected-Maximization (EM) algorithm. The EM algorithm refines the parameters of the GMM iteratively to increase the likelihood of the estimated model for the feature vectors being observed. According to Reynolds.

D. Speaker Model Adaption

The speaker-specific model is adapted from the UBM using the maximum a posteriori (MAP) estimation. The adaptation increases the performance and provides a tighter coupling between the two models

According to the alignment of the training vectors to the UBM can be computed as follows

$$pr\ (i \mid xt) = \frac{wi\ \ pi\ (xt)}{\sum_{j=1}^{m} wj\ pj\ (xt)}$$

III. RECOGNITION METHODOLOGY AND SYSTEM DESIGN

Previous studies have shown the accuracy of MFCC under low noise conditions and the robustness of GFCC in noisy environments. It would be beneficial to incorporate the benefits of these two approaches, to reduce or eliminate their individual drawbacks.

A. speaker combined feature representation

The strategy we are proposing allows us to combine the feature vector of MFCC and GFCC and use PCA to reduce the feature dimension and remove correlations.

The front-end block diagram of the system is depicted on Figure 6-1. The system is subdivided into two different subsystems: MFCC and GFCC. Both systems will be running in parallel during the training and test phases. The output of these systems is aggregated and processed using statistical PCA.

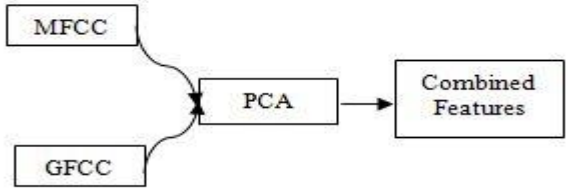

Fig 2: The combined feature representation front-end block diagram

These principal components are a linear combination of the optimally-weighted observed variables. These optimum basis vectors are the eigenvectors of the covariance matrix of the distribution.

B. Experimental setup

During the evaluation phase, each test segment is scored against the background model and a given speaker model to accept/reject the claim. The same set of tests is performed on both corpora. The experiment extracts 12-dimensional MFCCs from a pre-emphasized speech signal, mean and variance normalization and writes them to disk in HTK format. The second stage extracts 12-dimensional GFCC's from the same speech signal and stores it to the disk in HTK format as well. The last stage uses the output of the MFCC and GFCC as the input to the PCA function. To complete the experiment, the following steps are executed: UBM training, MAP adaptation, scoring of the verification trials, and computing the performance measures.

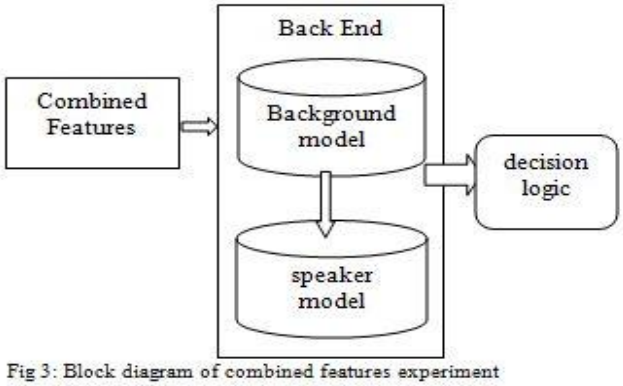

Fig 3: Block diagram of combined features experiment

IV. EVALUATION AND COMPARISON

The performance baselines used for comparison are the individual features, against the combined feature set. The Table I shows the summary of the EER achieved for each feature extraction technique

TABLE I : summary of the EER achieved for each feature extraction technique

SNR(dB)	EER % MFCC	EER % GFCC	EER % Combined	DCF
-30	49.929	25	25.303	10
-15	38.859	24	22.848	9.97
-10	27	22.879	17.121	9.49
-5	16	17.252	13.818	7.19
0	12	13.33	10.475	6.27

An important finding in our study is that GFCC features outperform conventional MFCC features under noisy conditions. By carefully examining all the differences between them, we conclude that nonlinear rectification mainly accounts for noise robustness differences. In particular, cubic root rectification provides more robustness to features than log rectification. In a noisy mixture, there are target dominant T-F units or segments indicative of this energy information. The cubic root operation makes features scale-variant (i.e. energy level dependent) and helps to preserve this information. The log operation, on the other hand, does not encode this information. Since MFCC is widely used in automatic speaker and speech recognition. We have conducted an in-depth study on the noise robustness of GFCC and MFCC features. Our experiments first confirm the superior robustness of GFCC relative to MFCC exists on a new corpus. By carefully examining all the differences between them, we conclude that the nonlinear rectification mainly accounts for the noise robustness differences. In particular, the cubic root rectification provides more robustness to the features than the log. In a noisy mixture, there are target dominant T-F units or segments indicative of this energy information. The cubic root operation makes features scalevariant (i.e. energy level dependent) and helps to preserve this information. The log operation, on the other hand, does not encode this information. Although the combined system in this chapter significantly outperforms the individual modules .The simple combination strategy in seems to lose its advantage when the performance profiles of the individual modules are similar. In such situations, more sophisticated methods of classifier combination may be needed.

V. CONCLUSION

A combined approach for feature extraction has been presented and compared with MFCC and GFCC feature extractions algorithms. The proposed combination feature methodology has shown satisfactory versatility and robustness under ELSDSR dataset. The final results in Table 10-1 shows that for the SNR levels tested overall there were significant improvement against the single feature counterparts. The highest improvement against MFCC was found at the -30dB range in which the EER improved 49%. The results also show that the combined MFCC-GFCC is indeed a viable method to improve recognition rates at low SNR levels

REFERENCES

[1] Bo Li, Khe Chai Sim,"A Spectral Masking Approach to Noise-Robust Speech Recognition Using Deep Neural Networks" IEEE/ACM Transactions On Audio, Speech, And Language Processing, VOL. 22, NO.8, AUGUST 2014

[2] Ji Ming, Timothy J. Hazen,James R. Glass and Douglas A. Reynolds,"Robust Speaker Recognition in Noisy Conditions," IEEE Transactions On Audio, Speech, And Language Processing, VOL. 15, NO. 5, JULY 2007

[3] Michael I. Mandel,Scott Bressler, Barbara Shinn-Cunningham, and Daniel P. W. Ellis,,"Evaluating Source Separation Algorithms With Reverberant Speech," IEEE Transactions On Audio, Speech, And Language Processing, VOL. 18, NO. 7, SEPTEMBER 2010

[4] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," Ieee Transactions On Audio, Speech, And Language Processing, VOL. 19, NO. 4, MAY 2011

[5] Ning Wang,P. C. Ching,Nengheng Zheng, and Tan Lee,"Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features," IEEE Transactions On Audio, Speech, And Language Processing, VOL. 19, NO. 1, JANUARY 2011.

[6] Tiago H. Falk and Wai-Yip Chan"Modulation Spectral Features for Robust Far-Field Speaker Identification," IEEE Transactions On Audio, Speech, And Language Processing, VOL. 18, NO. 1, JANUARY 2010

[7] Tobias May, Steven van de Par, and Armin Kohlrausch,"Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling," IEEE Transactions On Audio, Speech, And Language Processing, VOL. 20, NO. 1, JANUARY 2012

[8] William Hartmann, Arun Narayanan, Eric Fosler-Lussier, and DeLiangWang, "A Direct Masking Approach to Robust ASR," IEEE Transactions On Audio, Speech, And Language Processing, VOL. 21, NO. 10, OCTOBER 2013

[9] Xiaojia Zhao, Yuxuan Wang, and DeLiang Wang 'Robust Speaker Identification in Noisy And Reverberant Conditions' IEEE/ACM Transactions On Audio, Speech, And Language Processing, VOL. 22, NO. 4, APRIL 2014

[10] Yuxuan Wang, Kun Han, and DeLiang Wang,"ExploringMonaural Features for Classification-Based Speech Segregation," IEEE Transactions On Audio, Speech, And Language Processing, VOL. 21, NO. 2, FEBRUARY 2013

[11] Yuxuan Wang and DeLiang Wang"Towards

Scaling Up Classification-Based Speech Separation," EEE Transactions On Audio, Speech, And Language Processing, VOL. 21, NO. 7, JULY 2013

[12] Zhaozhang Jin, and DeLiang Wang,"Reverberant Speech Segregation Based on Multipitch Tracking and Classification," IEEE Transactions On Audio, Speech, And Language Processing, VOL. 19, NO. 8, NOVEMBER 2011

