Short-Term Spectral Feature Extraction and Their Fusion in Text Independent Speaker Recognition: A Review

Ruchi Chaudhary

Submitted in March, 2013; Accepted in August, 2013

Abstract - The paper gives an overview of Text-independent short-term-feature-extraction methods of Speaker Recognition System, for clean as well as noisy environment and their fusion at different levels. The basics of extracting feature, which is an imperative component for speaker recognition system, have been discussed along with their variants. The evolution of the conventional methods to 'State-of-the-Art' feature extraction methods are also brought out. This review helps in understanding the developments, which have taken place at various stages, with their relative merits and demerits. A comparative study of different techniques has been done at the end of each section to justify the choice of techniques available in the 'State-of-the-Art' speaker recognition systems. This study quantifies the effectiveness of short-term features for speaker identification.

Index Terms - Short-term feature Extraction, Speaker Recognition, Mel-frequency-Cepstral-coefficients (MFCC), Fusion.

1. INTRODUCTION

Speaker Recognition is the process of automatically recognizing the speaker by the use of the vocal characteristics [1-4]. 'State-of-the-Art' speaker recognition system uses number of voice characteristics, which include physical difference of the vocal production organs (shape of vocal tract, larynx size), and the manner of speaking (accent, rhythm, annotation style, pronunciation, pattern choice, vocabulary etc.)[1-4]. Fundamentally, speaker recognition process involves, extraction of speaker's specific characteristics (called features) from the given speech samples (*process known as feature extraction*) and the speaker model is trained and stored into the system database. In the recognition mode, the feature vector is extracted from the unknown's person's utterance and compared against the trained model. The purpose of feature extraction stage is to extract the speaker-specific information called feature vectors, represent the speaker-specific information based on one or more of the following: vocal tract, excitation source and behavioral tracts. All speaker recognition systems use set of scores to enhance the probability and reliability of the recognizer.

National Technical Research Organisation,
Old JNU Campus, Block-III, New Delhi-110067.
E-mail: ruchimakani@rediffmail.com

Speaker recognition systems can be divided into text-dependent and text-independent recognition systems [1-4]. In text-dependent speaker recognition systems, the speaker uses the same phrase at the time of enrolment/verification and in text-independent speaker recognition systems, which are more challenging and complex, the text of the speech at the time of enrolment/verification are completely random in nature. Success, in both recognition tasks, depends on extracting and modeling the speaker specific characteristics of the speech signal.

Pre-processing plays a vital role in speaker recognition as it decreases the acoustic effect and channel-mismatch. It is considered good practice to reduce the amount of variation in the data that does not carry important speech information. In other words, the pre-processing removes all non-relevant information such as background noise, characteristics of recording device etc. Voice activity detection (VAD), pre-emphasis filtering, normalization and mean subtraction are the few widespread commonly used steps in pre-processing [1-4]. By applying a pre-emphasis filter the glottal waveform and lip radiation characteristics are eliminated. The following fig. 1 shows the basic block diagram of speaker recognition system.

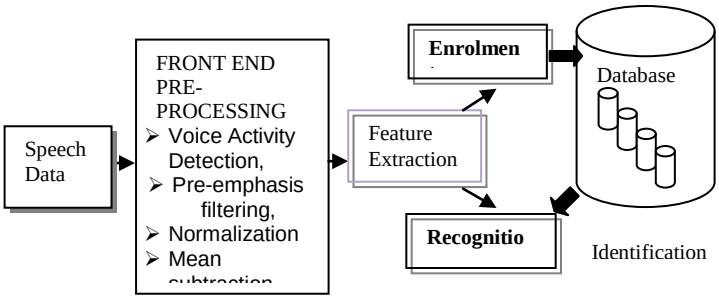

Figure 1: Block diagram of Speaker Recognition System

In order to have a good recognition performance, the front-end of the recognizer should provide feature vectors that can capture the important characteristics of an utterance. Besides, the front-end should also demonstrate reasonable performance in robust environment. Features can be categorized into *(1) Short-term spectral features* (2) *Voice source features* (3) *Spectro-temporal features* (4) *Prosodic and high level features*. Short-term spectral features are also referred as low-level features, have been dominantly used for speaker identification, as they are easy to compute and yield reasonably good performance, because these reflect information about the speaker's physiology and do not rely on the phonetic content

(which makes them inherently text-independent) [2]. Short-term analysis has been effective because of the quasi-stationary property of speech. The higher-level features also have the potential of increased robustness to channel variation, since phonetic, prosodic, and lexical information usage or temporal patterns do not change with the change of acoustic conditions [2]. Long-term information refers to features that are extracted over a longer region than a short-term spectral feature frame. Prosodic features capture variations in intonation, timing, and loudness that are specific to the speaker [2]. Because such features are supra-segmental i.e., extend beyond one segment, they can be considered a subset of long-term features.

The fundamentals of short-term features extraction methods used for speaker recognition systems, in clean and noisy environment, are discussed in section II and the overview of the fusion methods used in short-term-spectral-feature-based speaker identification system is provided in section III. Finally, section IV presents the conclusions of the study.

2. SHORT-TERM FEATURES EXTRACTOR

Generally, in any short–term feature extraction technique, framing and windowing are required before the mathematical computation of feature vectors. The general block diagram of short-term feature extraction is shown in the fig. 2.

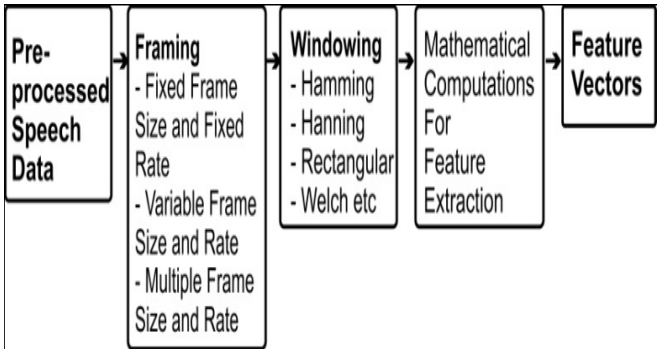

Figure 2: Block diagram of Short term Feature extraction

2.1 Framing

In short-term spectral feature (frame-based processing) are computed from frames size of about 20-30 ms in duration. Within this interval, the signal is assumed to remain stationary and a spectral feature vector is extracted from each frame. Generally, the frame shift is fixed to about half of the frame length. The *Fixed Frame size and Rate (FFFR)* approach results into some problems such as accidentally meeting noise frames. It may not be the best choice because characteristics of speech signal may rapidly change, especially at phonetic boundaries. The recognition accuracy increases, if the frame interval is directly controlled using phonetic information [5-11]. To avoid accidentally meeting noise frames problems, *Variable Frame Size and Rate (VFSR)* technique was proposed by *Qi eng Zhu et. al.* [5], which select optional frame size and frame rate depending on speaker rate, to capture sudden changes in spectral information with time. *Ponting et. al.* [6] and *Samkwong et. al.* [8] showed that VFSR could successfully improve the performance of speech recognition; however, it increases the burden in identifying the spectral changes in speech. The study in [9] demonstrated that the spectral variations in speech can also be captured by combining multiple frame size (MFS) and multiple frame rate (MFR). It was shown that combined MFS and MFR gives better performance, compared to fixed frame size for language and speech identification task. *Multiple frame size and rate (MFSR)*, proposed by *H S Jayanna et. al.* [9] shows that MFSR framing method increases the performance of speech recognition, as it can generate more number of feature vectors, in case of limited training/testing speech data. *Chi-Sang Jung et. al.*[10], in 2010, based on experiments, have shown that feature frame selection methods result, in minimum redundancy within selected frames, and maximizes the relevancy to speaker recognition characteristics which produces consistent improvement.

2.2 Windowing

The window functions such as rectangular, hamming, hanning, welch etc, which minimize the spectral distortion, are needed because of the finite-length effect of the Discrete Fourier Transform (DFT). An overlap between the windows is needed since there is loss of information at the borders of window. In speech processing a hamming window is mostly used as it gives more precise frequency estimation [12]. The global shape of the DFT magnitude spectrum known as spectral envelope contains information about the resonance properties of the vocal tract which is highly relevant for speaker identification. A simple model of spectral envelope uses a set of band-pass filters to do energy integration over neighbouring frequency band. In 1969, Fast Fourier Transform (FFT) which is a fast implementation of DFT, based cepstral coefficients, was used in study [13].

2.3 Feature Mathematical Computation Methods

To develop robust speaker identification system, it is necessary to understand the different feature mathematical computation methods for extracting the features from the speaker's speech. A good feature set should represent all the components of speaker information, however, there does not yet exist globally 'best' feature but the choice is a trade-off between speaker discrimination, robustness, and practicality.

S Pruzansky, *et. al.* [14-15], conducted the first speaker identification study in 1963. In his study he had shown that spectral energy patterns and their variance yielded good performance for the speaker recognition. Glenn [16] in his study, in 1967, suggested that acoustic parameters produced during nasal phonation are highly effective for speaker recognition and average power spectral of nasal phonation can be used as the features for the speaker recognition. In 1969, Fast Fourier Transform (FFT) based cepstral coefficients were used in speaker verification study [4].

In 1969, Atal [17] demonstrated the use of variations in pitch as a feature for speaker recognition. Other acoustic parameters

such as glottal source spectrum slope, word duration and voice onset were proposed as features for speaker recognition by Wolf [18] in 1972. The concept of Linear Prediction (LP) for speaker recognition was introduced by Atal in 1974 [19]. The basic idea behind linear predictive analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples [4]. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and linear predicted values, a unique set of parameters or predictor coefficients can be determined [4].

In 1976, *Sambur et. al*. [20-21] proposed the use of orthogonal linear prediction coefficients (LPC) as feature in speaker identification. LPC analyze the speech signal by estimating the formants, removing their effect from speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called residue [21]. In 1977, long term parameter averaging, which includes pitch, gain and reflection coefficients for speaker recognition was studied [22] and, it was shown that reflection coefficients are informative and effective for speaker recognition system. *Reflection Coefficients*, defined as a sequence of ratios of the discontinuity of the cross-sectional area of the vocal tract, various improved derivates of LPC i.e. Line Spectral frequency (LSF) [23], log Area Ratio (LAR) [24], Perceptual log Area Ratio (PLAR) [24], Perceptual Linear Prediction (PLP) [25] etc., were also studied by the researchers. Perceptual Linear Prediction (PLP) works by warping the frequency and spectral magnitudes of the speech signal based on auditory perception tests.

Mermelstein [26] conducted experiments in 1976, and showed that the cepstral coefficients are useful for representing consonantal information. Cepstral coefficients are the result of a cosine transform of the real algorithm of the short- term energy spectrum [27]. A study carried by *Rosenberg et. al.* [20] suggested that adjacent cepstral coefficients are highly correlated and hence all coefficients may not be necessary for speaker recognition. The LPCs are converted to ceptstral coefficients using autocorrelation techniques on linear frequency scale, called *Linear Predictive Cepstral Coefficients* (LPCCs). *Mel frequency capstral coefficient* (MFCC) were introduced in 1980, which use mel-frequency scale, are popular features in speech processing [4, 31-34]. The advantage of mel-frequency scale is that it approximates the nonlinear frequency resolution of the human ear. As with any filter bank based speech analysis technique, an array of band pass filter is utilized to analyze the speech in different frequency bandwidths. In MFCC parameterization, the position of the band pass filter along with the linear frequency scale is mapped to Mel-scale according to equation (1):

$$f_{mel} = 2595 \log_{10}(1+ f/100) \quad (1)$$

Overlapping of triangular filters in low frequency region of the energy spectrum (upto 1 KHz) in MFCC, are closely spaced. While smaller number of less closely spaced triangular filters, are used to cover the high frequency zone, to weight the DCT of the speech so that the output is approximately of the same order as the energies of the filter bank signals. The experiments conducted by researchers, have shown that as the number of filters in the filter-bank increases, the identification rate of system increases [33-39]. In [41], Gaussian filters (GF) are also suggested, as they improved the system performance over the conventional triangular filter. GF provide much smoother transition from one sub-band to the other preserving most of the correlation between them. The means and variances of these GFs can be independently chosen in order to have control over the amount of overlap with neighbouring sub bands.

A study by *Reynolds* et. al. [28], in 1994, compared the different features like MFCCs, LPCCs, LPCs and Perceptual Linear Prediction Cepstral Coefficients (PLPCCs) for speaker recognition. They reported that among these features, MFCCs and LPCCs gave better performance than other features. At present, even though various alternative features are available, the MFCC seem to be difficult to beat in practice.

In 1981 *Furui* [35] introduced the concept of dynamic features, to track the temporal variability in feature vector, in order to improve the speaker recognition performance. In addition to short-term frame energy, formant transitions and energy modulations also contain useful speaker-specific information. A common way to incorporate some temporal information to features is through 1st and 2nd order time derivative estimates, known as delta (Δ) and double delta (Δ^2) coefficients, respectively. A study made by *G R Doddington* in 1985 [36], converts the speech directly in to pitch, intensity and formant frequency and these features were also demonstrated to provide good performance.

Md. Sahidullah, et. al. [37], have proposed inverted filter bank structure, such that the higher frequency range is averaged by more accurately spaced filters and a smaller number of widely spaced filters are used in the lower frequency range. This feature set named as *Inverted Mel Frequency Cepstral Coefficients (IMFCC)* follow the same procedure as normal MFCC, but using reversed filter bank structure. The combination of both enhances accuracy [38-39] and as such they can be considered complementary.

Another approach using Wavelet Transform instead of Discrete Cosine Transform, in the feature extraction stage, was proposed by Nengheng Zheng, *et.al* [40] in 2007. According to this, WOCOR is derived by wavelet transformation of the LP residual signal and is capable of capturing the spectro-temporal properties of vocal source excitation. There has been an array of these features used such as wavelet filter banks. It has been shown that the speaker identification system outperform when combination of both MFCC and WOCOR are used as feature extraction.

2.4 Features Extraction Methods for Noisy Environment

Robust speech techniques attempt to maintain the performance of a speech processing system under diverse conditions of operation (environmental differences between training/testing

conditions). To improve the performance of speaker recognition systems in noisy environment, approaches can be roughly divided into three categories, namely, *robust speech feature extraction*, *speech enhancement* and *model-based compensation* for noise, as shown in fig 4.

In the case of *speech enhancement*, some initial information about speech and noise is needed to allow the estimation of noise and clean up of the noisy speech. Widely used methods in this category include spectral subtraction (SS) and Wiener filtering. Statistical models such as Hidden Markov Models (HMMs), *Parallel Model Combination* (PMC), *Vector Taylor Series* (VTS) and *Weighted Projection Measure* (WPM) are generally classified into model-based compensation category [41-46]. The compensation techniques try to remove the mismatch between the trained models and the noisy speech, to improve the performance of the system.

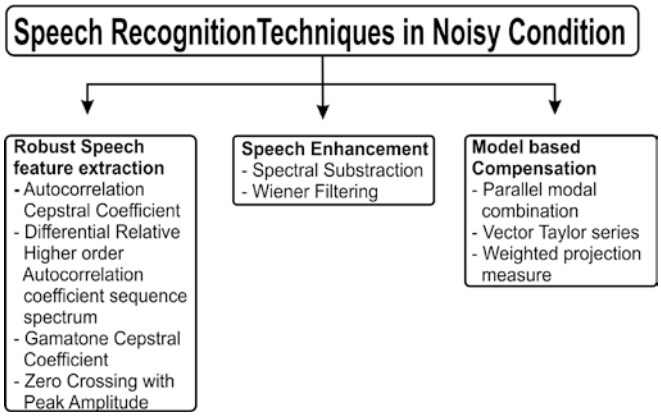

Figure 4: Speech Recognition techniques in noisy environment

LPCC, PLP and MFCC, perform well in clean environment, however, they suffer severe performance degradation in noisy conditions, especially when there is a noise level mismatch between the training and testing environments [41-46]. One of the major issue with these methods is that they are very sensitive to additive noise. Use of, *robust speech feature extraction* methods for improvement in speaker identification system in noisy environment is discussed in following paragraphs.

Use of the autocorrelation domain in speech feature extraction has recently proved to be successful for robust speech recognition [46]. A temporal filtering procedure on the autocorrelation sequence has been proposed [47] to minimize the effect of additive noise, which is derived based on filtering the temporal trajectories of short time one sided autocorrelation sequence. This filtering process minimizes the effect of additive noise which are stationary in nature at low SNR‟s, white, F16 and factory noise and no prior knowledge of noise characteristics is required. Results of experiments conducted at [47-48], indicate that *an Autocorrelation Mel Frequency cepstral coefficient (A-MFCCs)* significantly improves the performance of speaker identification system in noisy environment.

Poonam Bansal, *et. al.* [49] in 2010 evaluated the speech recognition performance of the *Differentiated Relative Higher Order Autocorrelation Coefficient Sequence Spectrum* (DRHOASS) features. DRHOASS uses only the higher-order autocorrelation coefficients for spectral estimation and discards the lower order autocorrelation coefficients. Speech coefficients, using the magnitude spectrum of the relative one-sided higher-order autocorrelation sequence, differentiating it and then processing it through a Mel filter bank, were finally parameterized in terms of MFCCs. It is shown in [49] that DRHOASS features perform almost similar to MFCC features, for clean speech; however, it performs better than the MFCC features for noisy speech. It was found that higher order autocorrelation coefficients along with additional filtering improved the robustness of the speech recognition system under different background noises.

Gammatone Cepstral Coefficients (GTCC) and *Zero-Crossings with Peak Amplitude (ZCPA)* are also claimed to have better performance than the conventional algorithms, especially in extremely noisy conditions (<15dB SNR) [50, 51]. *GTCC*, an acoustic feature extraction based on an auditory filterbank realized by Gammatone filters, was introduced for large vocabulary speech recognition [50]. Gammatone filters were used to characterize data obtained by reverse correlation from measurements of auditory nerve responses. The zero-crossing analysis (ZCA) of speech waveforms, proposed by *Doh-Suk Kim, et. al.* [51] has advantages over autocorrelation, power spectrum and linear prediction methods. This is because of the reason that these conventional methods data extraction by sampling a time waveform, depends on the maximum frequency content in the time signal whereas, ZCA requires a number of extracted samples, determined by the average rate of zero-crossing intervals. In ZCPA, the reciprocal of time intervals between two successive zero crossings are collected in frequency histograms from which frequency information is extracted.

In summary, spectral features like band energies, formants, spectrum and cepstral coefficients represent mainly the speaker specific information due to the vocal tract. Excitation source feature like pitch, variation in pitch information from LP residual and glottal source parameters represents mainly the speaker specific information due to vocal cord. Long-term, features like duration, intonation, energy, AM and FM components represents mainly the speaker specific information due to the behavioural traits. Practically use of MFCC & LPCC as short term feature extraction method for speaker specific information provides accuracy and reliability. However, GTCC, ZCPA, DRHOASS provide better performance in noisy environment.

3. FUSION OF COMPLEMENTARY INFORMATION

Recent researches have indicated that the appropriate fusion is an approach to meet stringent performance requirement of a recognition system. The fusion hold the promise of improving basic recognition accuracy by adding complementary information, not captured by the conventional features alone

and, possibly, robustness to acoustic degradations from channel and noise effects, to which majority of the features are highly susceptible. Fusion reduces the cost of implementation by using several cheap sensors rather than one expensive sensor [52-57, 60-61].

A common recognition system includes fingerprint, face, hand geometry, finger geometry, iris, retina, signature, voice, gait, smell, keystroke, ECG, etc. Recognition systems based on single source of information are called unimodel systems. Since, the unimodel system alone may not be able to achieve the desired performance requirement in real world applications, the use of multimodal biometric authentication system, which combines information from multiple modalities increases the performance [52]. The scope of this section is to cover the fusion preferred for unimodel short-term-feature-based speaker identification system in order to enhance the performance system.

The performance of the information fusion system is highly dependent on the effectiveness of the fusion technique implemented. By considering pre-defined data attributes, including channel characteristics and speaker's emotional and stress patterns detectable in conversations, the fusion method need to be fine-tuned to improve results. Over a period of time, numbers of information fusion techniques have been proposed for speaker identification systems. Generally, the information fusion can be done at: (i) feature level, (ii) score level, or (iii) decision level [55]. The block diagram of fused system (different level of fusion) is shown in fig 5. In subsequent paragraphs, fusion strategy and various fusion levels methods for unimodel short-term-feature-based speaker identification system are discussed and the comparison in term of their performance, is also brought out.

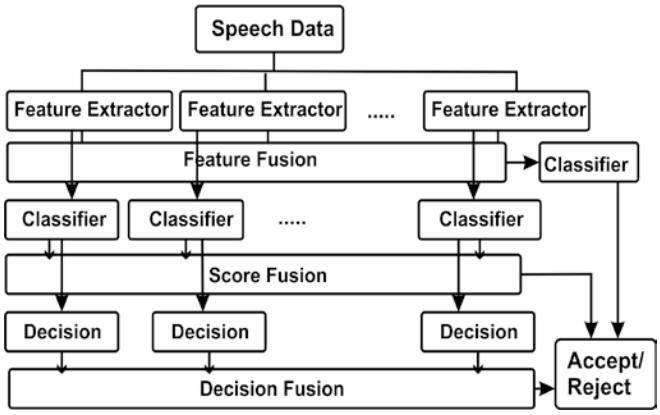

Figure 5: Block diagram of Fused System

3.1 Fusion strategy

In course of time, various architectures and schemes have been directed by researchers for combining multiple features/classifiers. Authors at [57-58] developed a common theoretical framework for combining features or classifiers which use distinct pattern representations. Number of possible combination schemes such a product, sum, minimum, maximum, median, average etc., have been utilized and their performance compared empirically. It is found that the sum rule outperforms the other combination schemes and it is most resilient to estimation errors [57-58].

3.2 Feature Level Fusion

In feature fusion, multiple features are concatenated into a large feature vector, and thereafter, a single model is trained with these fused large feature vector [59]. Each feature vector need to be normalized before concatenated. A well-known data fusion strategy in speaker identification system is to concatenate the cepstral vectors with their delta and delta-delta cepstral into a long feature vector. Also the fundamental frequency has been used in addition with the cepstral vectors to improve recognition accuracy. Researchers have used various accent features such as pitch, energy, intonation, MFCCs, formants, formant trajectories, etc., and some have fused several features to increase accuracy as well. In [59], a fusion of MFCC, accent-sensitive cepstral coefficients (ASCC), delta ASCCs, energy, delta energy and delta-delta energy was done to improve the accuracy of the speaker identification system. In commercial biometric systems, the MFCC, Delta Coefficients and formant feature fusion method are preferred over the other fusion techniques.

3.3 Score Level Fusion

In score level fusion, each different feature set is processed separately with specialized classifier, and thereafter, output scores obtained from each classifier are combined. Each of the different features sets acts as an independent "expert", giving its opinion about the unknown speaker's identity. The *fusion rule* then combines the individual experts match scores. This approach is referred to as *score-level fusion*. Score level fusion is generally preferred in systems since the matching scores relatively easy to obtain and contain sufficient information to make genuine and impostor case distinguishable. The score mean and variance of multiple non-target speaker models are used to normalize the score of the target speaker models. Since the scores generated by identification system can be either similarity scores or distance scores, one needs to convert these scores into a same nature. The common practice, which is followed, is to convert all the scores into similarity scores.

Researchers at [37-39], used MFCC and IMFCC feature vectors, with two separate Gaussian Mixture Model (GMM) classifiers and fused their scores. Likewise, same principle has been adopted for Gaussian filter based MFCC and IMFCC also [37-39]. In both cases, for each speaker, two scores were generated; improved system performances were observed after fusion in both the cases. In [59], two support vector machine (SVM) classifiers, using both MFCCs and LPCCs separately, were fused to achieve high accuracy of the system.

In 1995 *P. Thevenaz et. al.* [62] reported that the energy of LP residual alone, gives less performance and combining it with LPCC improves the performance as compared to that of LPCC alone. Several other studies have been reported that combining

LP residual with MFCC improves the performance as compared to that of MFCC alone [63]. In 1996, *Plumpe* developed a technique for estimating and modeling the glottal flow derivative waveform from speech for speaker recognition [64]. In this study, the glottal flow estimate was modelled as coarse and fine glottal features, which were captured using different techniques. Also it was shown that combined coarse and fine structure parameters gave better performance than the individual parameter alone. In 2003, *B Peskin, et. al.* [65] reported that combination of prosodic features like long-term pitch with spectral features provided significant improvement as compared to only pitch features. *Nengheng Zheng, et. al.* [40] in 2007 studied that the complementary contributions of MFCC and WOCOR in speaker identification, significantly outperforming the one using MFCC only. Thereafter, a score level fusion technique was used for combining MFCC and WOCOR for speaker identification, in order to improve the performance. Further, the system comprised of a fusion of classification scores from adapted HMM and GMM, where scores from two recognition systems were fused [66].

The area of automatic speaker recognition has been dominated by systems using only short-term, such as cepstral features. These systems generally produced low error rates. Recently published works have demonstrated that such high-level information can also be used successfully in automatic speaker recognition systems by improving accuracy and potentially increasing robustness. Wide ranging high-level-feature-based approaches using pronunciation models, prosodic dynamics, pitch gestures, phone streams, and conversational interactions were explored. In [66], vocal tract and vocal cord parameters are integrated to identifying speakers. In spite of the two approaches having significant performance differences, the way they represent speech signal is complementary to one another. In [67-68] have also shown great improvement in speaker verification accuracy through fusion of low and high speech levels classifiers. However, the complexity of the system increases.

3.4 Decision Level fusion

Each of the expert classifier produces an identified speaker label, and the fusion rule combines the individual decisions e.g. by majority voting, called *decision level fusion* strategy. In other words, it is hardening the decisions of the individual classifiers. The inclusion of multi-resolution classifiers enhances fusion capabilities in handling noisy patterns, thus increasing accuracy. Other approaches to combine classifiers (decision fusion) include the rank-based methods such as the Borda count, the Bayes approach, the Dempster-Shafer theory, the fuzzy integral, fuzzy connectives, fuzzy templates, probabilistic schemes, and combination by neural networks [69-76].

The results at [61] show that the four acoustic feature based subsystems outperform the tokenization subsystems and the best subsystem is the LPCC-SVM system. Comparisons with the MFCC-GMM system, the temporal discrete cosine transform (TDCT)-GMM system captures the longer time dynamic of the spectral features but it requires more training data.

Author at [76] has shown that the decision fusion based on the output of the GMM and SVM classifiers increases the discriminative power, as does fusion between classifiers based on spectral features and classifiers based on prosodic information.

3.5 Comparison on Fusion Methods

Although feature-level fusion may improve recognition accuracy, it has several short comings. Firstly, fusion becomes difficult if a feature is missing (e.g. F0 of unvoiced sounds) or the frame rates of the features are different. Secondly, the number of training vectors needed for robust density estimation increases exponentially with the dimensionality. This phenomenon is known as the *curse of dimensionality*. It can be seen that the feature-level fusion improves the performance over the individual classifier in the case of MFCC and its delta features. However, in all other cases it degrades the performance [77]. Decision level fusion is considered to be rigid due to the availability of limited information. However, this is the best fusion strategy, when all feature sets are used. Features using parametric linear prediction based estimation of the spectral envelope (LPCC variants) provide the best speaker recognition results and are the most useful in fusion, followed by minimum variance distortion-less response based estimation. The score level fusion gives the best results in all cases fusing feature. It can be seen that the feature-level fusion improves the performance over the individual classifier in the case of MFCC and its delta features. Thus, fusion at the score level is usually preferred, as it is relatively easy to access and combine the scores presented by the different modalities. In [78], score and decision fusion for the mel-cepstrum and corresponding delta features and demonstrated that the score level fusion performed consistently better. Furthermore, they observed that the computational complexity for the feature fusion is higher than that of the score and decision level fusion.

Hybrid feature based speaker identification system has been proposed in [79], and various combination of the feature-level, score-level and decision-level has been observed to give the advantage on over the single level fusion methods.

For any text independent short-term-feature-based speaker identification system, fusion at the score level is generally considered appropriate due to the ease in accessing and consolidating matching scores.

4. CONCLUSIONS

In this paper, short-term feature extractions techniques are discussed for text-independent speaker recognition system in clean as well as in noisy environment and also including various fusions preferred in order to maximize system performance for speaker identification system.

Among the developed techniques, the speaker recognition system widely uses short-term features for speech analysis; MFCC and its derivatives as robust feature extraction techniques. The combination of various variant of MFCC and

the other feature extraction techniques can be considered according to the intended application. In summary, score level fusion of short term features, mainly LPCC & MFCC, based speaker identification system performs well as compared to other short term feature extraction methods and fusion levels.

ACKNOWLEDGEMENT

The author acknowledge the help provided in carrying out this study by Dr. M. S. Vijayaraghawan, Air(Cmdr) V. Sehgal, Shri Manuj Modi and the author is also grateful to Chairman, NTRO for providing infrastructural support, and necessary permission for the study.

REFERENCES

[1] J. Campbell, "Speaker recognition: a tutorial", *Proceedings of the IEEE*, vol.85, issue 9, pp. 1437–1462, Sep 1997.

[2] Tomi Kinnunen, Haizhou Lib, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors", *Speech Communication*, vol. 52, Issue 1, pp. 12–40, Jan 2010.

[3] H. A. Murthy, F. Beaufays, L. P. Heck and M. Weintraub, "Robust Text-Independent Speaker identification over Telephone Channels", *IEEE Transaction on Speech and Audio Processing*, vol. 7, No. 5, Sep 1990.

[4] H. S. Jayanna, S. R. Mahadeva Prasanna, "Analysis, Feature Extraction, Modelling and Testing Techniques for Speaker Recognition", *IETE Technical Review*, vol. 3, Issue 3, May-June 2009.

[5] Q. Zhu, A. Alwan, "On the use of variable Frame Rate Analysis in Speech recognition", *International Conf. on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1783-1786, 2000.

[6] K. M. Ponting, S. M. Peeling, "Experiments in Variable Frame Rate Analysis for Speech Recognition", Defense Technical Information Centre OAI-PMH Repository (United States), Dec 1989.

[7] D. Pekar, N. Jakovljevic, M. Janev, D. Miskovic, V. Delic, "On the use of Higher Frame rate in the Training Phase of ASR", ACM Digital Library, *Pro. of the 14th WSEAS International conf. on Computers*, vol. 1, pp. 127-130, 2010.

[8] Sam Kwong, Qian-Hua He, "The Use of Adaptive Frame for Speech Recognition", *Journal on Applied Signal Processing*, vol 2, pp. 82–88, 2001.

[9] H. S. Jayanna, S. R. Mahadeva Prasanna, "Limited data speaker identification", *Indian Academy of Sciences*, vol. 35, Part 5, pp. 525–546, Oct 2010.

[10] Chi-Sang Jung, Moo Young Kim, Hong-Goo Kang, "Selecting Feature Frames for Automatic Speaker Recognition Using Mutual Information", *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, Issue 6, pp. 1332 – 1340, Aug. 2010.

[11] R. Pawar, H. Kulkarni, "Analysis of FFSR, VFSR, MFSR Techniques for feature Extraction in Speaker Recognition: A Review", *International Journal of Computer Science*, vol. 7, Issue 4, No 1, July 2010.

[12] A. Oppenheim, R. Schafer, J. Buck, "*Discrete-Time Signal Processing"*, second ed. Prentice Hall, 1999.

[13] J. E. Luck, "Automatic Speaker verification using Cepstral Measurements", *Journal of the Acoustical Society of America*, Vol. 46, Issue 4B, pp. 1026-1032, Nov 1969.

[14] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition", *Journal of the Acoustical Society of America*, vol. 35, Issue 3, pp. 354-358, 1963.

[15] S. Pruzansky, M. V. Mathews, P. B. Britner, "Talker-Recognition Procedure Based on Analysis of Variance", *Journal of the Acoustical Society of America*, Vol. 35, Issue 11, pp. 1877-1877, 1963.

[16] J. W. Glenn, N. Kleiner, "Speaker Identification Based on Nasal Phonation", *Journal of the Acoustical Society of America*, vol. 43, Issue 2, pp. 368-372, 1968.

[17] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours", *Journal of the Acoustical Society of America*, vol. 45, Issue 1, pp. 309-309, 1969.

[18] J. J. Wolf, "Efficient Acoustic Parameters for Speaker Recognition", *Journal of the Acoustical Society of America*, vol 51, Issue 6B, pp. 2044-2056, 1972.

[19] B. S. Atal, "Effectiveness of Linear Perdition characteristics of the speech wave for automatic speaker identification and verification", *Journal of the Acoustical Society of America*, vol. 55, pp. 1304-1312, 1974.

[20] A. Rosenberg, M. Sambur, "New techniques for automatic speaker verification", *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 23, Issue. 2, Apr 1975.

[21] M. Sambur, "Selection of acoustic features for speaker identification", *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 23, Issue. 2, Apr 1975.

[22] J. Markel, B. Oshika, A. Gray, "Long-term feature averaging for speaker Recognition", *IEEE Trans. on Acoustic Speech and Signal Processing*, Vol. 25, Issue 4, pp. 330 - 337, Aug 1977.

[23] F. Soong, B. Juang, "Line spectrum pair (LSP) and speech data compression", *IEEE International Conf. on Acoustics Speech and Signal Processing*, pp 37–40, Mar 1984.

[24] D. Chow, W. Abdulla, "Robust Speaker Identification Based on Perceptual Log Area Ratio and Gaussian Mixture Models", *8th International Conf. on Spoken Language Processing*, vol. III, pp. 1761-1764, *2004*.

[25] F. zohra Chelali, A. Djeradi, R. Djerad, "Speaker Identification System based on PLP Coefficients and Artificial Neural Network", *Proceedings of the World Congress on Engineering*, London, U.K., vol II , July 2011.

[26] P. Mermelstein, "Distance Measures for Speech Measurements – Psychological and Instrumental", *Joint workshop on Pattern Recognition and Artificial Intelligence*, Hyannis, Mass, June 1976.

[27] S. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transactions on Acoustics Speech and Signal Processing*, vol 28 Issue 4, pp. 357 – 366, Aug 1980.

[28] D. A. Reynolds, R. C. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture speaker model", *IEEE Transaction on speech and Audio processing*, vol. 3, No. 1, Jan 1995.

[29] M. R. Hasan, M. Jamil, MGRMS. Rahman,"Speaker Identification using Mel frequency Cepstral Cefficients", *3rd International Conf. on Electrical & Computer Engg*, Dhaka, Bangladesh, pp. 28-30, Dec 2004.

[30] F. Zheng, G. Zhang, Z. Song, "Comparison on Different implementations of MFCC", *Journal of Computer Science & Technology*, vol.16, issue 6, pp. 582-589, Sept. 2001.

[31] E. Ambikairajah, "Emerging features for speaker recognition", *6th International IEEE Conf. on Information, Communications & Signal Processing*, Singapore, pp. 1–7, Dec 2007.

[32] Qi Li, Y. Huang, "Robust Speaker identification Using an Auditory Based Feature", *IEEE Int. Conf. on Acoustics Speech and Signal Processing*, pp. 4514-4517, Mar 2010.

[33] A. Mousa, "Mare Text Independent Speaker Identification based on K-mean Algorithm", *International Journal on Electrical Engineering and Informatics*, vol 3, No. 1, pp. 100-108, 2011.

[34] A. Lawson, P. Vabishchevich, M. Huggins, P. Ardis, B. Battles, A. Stauffer, "Survey and evaluation of acoustic features for speaker recognition", *IEEE International Conf. on Acoustics Speech and Signal Processing*, pp. 5444-5447, 2011.

[35] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features", *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 29, Issue 3, pp. 342 – 350, Jun 1981.

[36] G. R. Doddington, "Speaker Recognition–Identifying people by their Voices", *Proceedings of IEEE*, vol 73, pp. 1651-1664, Nov 1985.

[37] M. Sahidullah, S. Chakroborty, G. Saha, "Improving Performance of Speaker Identification System Using Complementary Information Fusion", *Proceedings of 17th International Conference on Advanced Computing and Communications*, pp. 182-187, 2009.

[38] S. Chakroborty, G. Saha, "Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter", *International Journal of Signal Processing*, vol 5, No. 1, pp 11-19, 2009.

[39] S. Singh, E. G. Rajan, "Vector Quantization approach for Speaker Recognition using MFCC and Inverted MFCC", *International Journal of Computer Applications*, vol 17, No. 1, pp. 0975- 8887, March 2011.

[40] N. Zheng, T. Lee, N. Wang, P. C. Ching, "Integrating Complementary Features from Vocal Source and Vocal Tract for Speaker Identification" *Computational Linguistics and Chinese Language Processing*, Vol. 12, No. 3, pp. 273-290, Sep 2007.

[41] M. Grimn, K. Roschel "*Robust Speech Recognition and Understanding*", InTech , Austria, June 2007.

[42] J. Ming, T. J. Harzen, J. R. Glass, D. A. Reynolds, "Robust Speaker recognition in Noisy Conditions", *IEEE Transactions on Audio speech and language processing*, vol. 15, No. 5, July 2007.

[43] D. S. Kim, S. Y. Lee, R. M. Kil, "Auditory Processing of speech signals for robust speech recognition in real world noisy environments", *IEEE Trans. Speech Audio Process*, vol. 7, no. 1, pp 55-69, Jan. 1999.

[44] Y. Gong**, "**Speech recognition in noisy environments: A survey", *Speech Communication*, vol. 16, Issue 3, April 1995, pp. 261-291, Dec 1999.

[45] B. J. Shannon, K. K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition", *Speech Communication*, vol. 48, pp. 1458–1485, Aug 2006.

[46] A. Dev, B. Parmanand, "A Novel Feature Extraction Technique for Speaker Identification", *International Journal of Computer Applications*, vol. 16, No.6, pp. 0975–8887, Feb 2011.

[47] T. E. Bachir, A. Benabbou, M. Harti, "Design of an automatic Speaker Recognition System based on Adapted MFCC and GMM methods for Arabic speech", *International Journal of computer Science and network Security*, vol. 10, No. 1, Jan 2010.

[48] S. Kim, M. Ji, H. Kim, "Robust speaker recognition based on filtering in autocorrelation domain and sub-band feature recombination", *Pattern Recognition Letters*, vol. 3, pp. 593–599, 2010.

[49] P. Bansal, A. De, S. B. Jain, "Robust Feature Vector Set Using Higher Order Autocorrelation Coefficients", *International Journal of Cognitive Informatics and Natural Intelligence*, vol. 4, issue 4, pp. 37-46, October-December 2010.

[50] R. Schluter, L. Bezrukov, H. Wagner, H. Ney, "Gammatone Features and Feature Combination for Large Vocabulary Speech Recognition", *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. IV-649 - IV-652, June 2007.

[51] D. S. Kim, J. H. Jeong, J. W. Kim, S. Y. Lee, "Feature extraction based on zero-crossings with peak amplitudes for robust speech recognition in noisy environments", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 61 – 64, vol. 1, 7-10 May 1996.

[52] D. L. Hall, J. Llinas, "*Handbook on Multi-sensor Data Fusion*", CRC press, Edition 1, June 2001.

[53] F. Huenupan, N. B. Yoma, C. M., C. Garreton, "Confidence based multiple classifier fusion in speaker verification", *Journal Pattern Recognition Letters, ACM Digital Library*, vol 29 Issue 7, pp. 957-966, May 2008.

[54] J. Keshet, S. Bengio, "*Automatic Speech and Speaker Recognition: Large margin and Kernel methods*", John Wiely & Sons Ltd, 2009.

[55] A. Ross, A. Jain, "Information fusion in biometrics", *Elsevier Science Pattern Recognition Letters*, vol. 24, pp. 2115–2125, 2003.

[56] J. Kittler, M. Hatef, R. Duin, and J.Matas. "On combining classifiers", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 20, issue 3, pp. 226–239, March 1998.

[57] J. Kittler, "Combining Classifiers: A Theoretical Framework", *Pattern Analysis & Applied. Springer-Verlag London Limited*, Issue 1, pp.18-27, 1998.

[58] J. Kittler, F.M. Alkoot, "Sum Versus Vote Fusion in Multiple Classifier Systems", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25 Issue: 1, pp. 110 – 115, Jan 2003.

[59] K. Chen, L. Wang, H. Chi, "Methods of Combining multiple Classifiers with Different Features and Their Applications to Text-Independent Speaker Identification", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 11, issue 3, pp. 417-445, 1997.

[60] K. R. Farrell, R. J. Mammone, K. T. Assaleh," Speaker Recognition using Neural Networks and conventional classifiers", *IEEE transaction on speech and audio processing*, Vol. 2, No. 1, Part II, Jan 1994.

[61] R. Tong, B. Ma, K. A. Lee, C. You, D. Zhu, T. Kinnunen, H. Sun, M. Dong, E. S. Chng, H. Li, "The IIR NIST 2006 Speaker Recognition System: Fusion of Acoustic and Tokenization Features" *Proceedings of the 5th International conf. on Chinese Spoken Language Processing, ACM digital Library*, pp. 566-577, 2006.

[62] P. Thevenaz, H Hugli, "Usefulness of the LPC-residue in text-independent speaker verification", *Speech Communication archive*, vol. 17, Issue 1-2 Aug 1995.

[63] S. R. M. Prasanna, C.S. Gupta, B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech", *Speech Communication*, vol. 48, pp. 1243-61, 2006.

[64] M. D. Plumpe, T.F. Quatieri, D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification", *IEEE Trans. Speech Audio Process.*, vol. 7, issue 5, pp. 569-85, 1999.

[65] B. Peskin, J. Navratil , J. Abramson, D. Jones, D. Klusacek , D. A. Reynolds, B. .Xiang, "Using prosodic and conservational feature for high-performance speaker recognition", *Int. Conf Acoustic, Speech, Signal Processing, Hong Kong*, vol. IV, pp. 784-7, Apr. 2003.

[66] J. P. Campbell, D. A. Reynolds, R. B. Dunn, "Fusing High- and Low-Level Features for Speaker Recognition", *Proc. European Conf. Speech Communication Technology*, pp.2665-2668, 2003.

[67] Y. A. Solewicz, M. Koppel, "Using Post-Classifiers to Enhance Fusion of Low and High-Level Speaker Recognition", *IEEE Trans. Audio Speech and Language Processing*, Vol. 15, No. 7, Sep 2007

[68] M. Salkhordeh, H. A. Vahedian, H. Sadoghi, Y. H. Modaghegh, "Designing Kernel Scheme for Classifiers Fusion**",** *International Journal of Computer Science and Information Security*, vol.6, No. 2, 2009.

[69] N. M. Wanas, R. A. Dara, M. S. Kamel, "Adaptive fusion and co-operative training for classifier ensembles", *Elsevier Pattern Recognition*, vol 39, pp.1781 – 1794, 2006.

[70] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, A. Gelzinis, "Soft combination of neural classifiers: A comparative study", *Elsevier Pattern Recognition Letters*, vol. 20, pp. 429-444, 1999.

[71] R. Tong, B. Ma, K. A. Lee, C. You, D. Zhu, T. Kinnunen, H. Sun, M. Dong, E. S. Chng, H. Li, "Fusion of Acoustic and Tokenization Features for Speaker Recognition", *Chinese Spoken Language Processing Lecture Notes in Computer Science*, vol 4274, pp 566-577, 2006.

[72] B. Reuderink, M. Poel, K. Truong, R. Poppe, M. Pantic, "Decision-Level Fusion for Audio-Visual Laughter Detection", *Springer-Verlag Berlin Heidelberg*, pp. 137–148, 2008.

[73] A. Mishra, "Multimodal Biometric it is: Need for Future Systems", *International Journal of Computer Applications*, vol. 3, No. 4, pp. 28- 33, June 2010.

[74] Z. Wu, L. Cai, H. Meng , "Multi-level Fusion of Audio and Visual Features for Speaker Identification" *Springer-Verlag Berlin Heidelberg*, pp. 493– 499, 2005.

[75] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. Leeuwen, P. Matejka, P. Schwartz, A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006", *IEEE Trans. Audio Speech and Language Processing* vol.15, No.7, pp. 2072–2084, Sep 2007.

[76] B. Reuderink, M. Poel, K. Truong, R. Poppe, M. Pantic, "Decision-Level Fusion for Audio-Visual Laughter Detection" *Machine Learning for Multimodal Interaction Lecture Notes in Computer Science Springer-Verlag Berlin Heidelberg*, vol 5237 , pp. 137–148, 2008.

[77] T. Kinnunen, V. Hautamäki, P. Fränti, "Fusion of Spectral Feature Sets for Accurate Speaker Identification", 9th Conference Speech and Computer, Saint-Petersburg, Russia Sep 20-22, 2004

[78] L. Valet, G. Mauris, P. Bolon, "A Statistical Overview of Recent Literature in Information Fusion" IEEE

Magazine on Aerospace and Electronic Systems, vol. 16, Issue: 3, pp. 7 – 14, 2001.

[79] M. R. Islam, M. F. Rahman, "Hybrid Feature and Decision Fusion based Audio-Visual Speaker Identification in Challenging Environment", *International Journal of Computer Applications*, vol. 9, No.5, Nov 2010.