THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

# A Survey on Robust Speaker Recognition for Noise and Channel Robustness

**J. V. Thomas Abraham**
Assistant Professor (Selection Grade), School of Computer Science and Engineering,
VIT University Chennai, Tamil Nadu, India

***Abstract:***
*Speaker Recognition is one among many research topics that researchers have done an umpteen research but still a popular research topic for many researchers. For more than three decades research has been carried out in this field but still many researches is going on in building a robust speaker recognition model. In this survey paper, the evolution of speaker recognition from its initial stage till the state-of-art technique is discussed in detail and the various techniques are evaluated based on their effectiveness.*

***Keywords:*** *Speaker recognition, MFCC, SVM, HMM, GMM, i-vector, PLDA*

## 1. Introduction

People express their thoughts in numerous ways and among them speech is the most convenient and effective way of communication. It is easy to gather the speech samples of people using simple devices such as telephones, microphones, any audio/video devices and hence it is the best choice of remote authentication. The most common method employed in authenticating a person is verifying the object he/she possess and then his knowledge (e.g. PIN and passwords). These can easily be stolen by an intruder or forgotten by the person himself. A more secured and sophisticated technique is to use the biometric data such as fingerprints, face and voice of a person which are unique to the individual and exemplify the individual. And these data cannot be forgotten or stolen by someone else. As speech is the easiest data to collect, persistent and less obtrusive, a system that processes the speech and recognizes the individual receives constant attention of the researchers. Speaker Recognition (SR) also commonly known as voice recognition is identifying the person to whom the voice belongs to rather than what was spoken. The latter is known as speech recognition. The first type of speaker recognition system in the 1960's uses spectrogram of voices, also known as voiceprint analysis. It is the acoustic spectrum of the voice that is similar to the fingerprint. However, this type of analysis could not fulfill the aim of automatic recognition as human interpretation was needed. Later in 1980's different types of features were extracted from speech. These features were represented in time, frequency or in both domains and used for speaker recognition.

Two broad divisions of Speaker recognition are Speaker Identification (SI) and Speaker Verification (SV). From the literature review one can find that the term speaker verification is also referred as voice verification, speaker authentication, voice authentication, and talker verification and talker authentication.

Speaker Identification is finding a match between the testing (unknown speaker's) speech signals with any of the (known speaker's) speech signal which has been modeled already. Speaker verification is checking the claim of a person about his identity is true or not. So speaker identification is 1: N matching process while speaker verification is 1:1 matching process. Speaker identification or speaker verification can be further classified as text-dependent or text-independent. In a text-dependent system, user is expected to say the same words for which the model has been created. In a text-independent system the user need not to speak the same words to recognize one.

Speaker identification can be further classified into open and closed set recognition. In open set recognition, the system is able to suggest that the voice from the unknown speaker does not match any speaker in the registered database. In closed set recognition, the voice will come only from the specified set of known speakers and the system is forced to make a decision based on the best matching speaker in the registered database.

This paper is organized as follows: Section 2 briefly explains the various tasks involved in any SRS. Section 3 gives an overview of the several basic SRS. Issues and challenges faced with these SRS have been discussed in Section 4. The robust SRS addressing the different issues is given in Section 5. The performances of the robust SRS are discussed in Section 6.

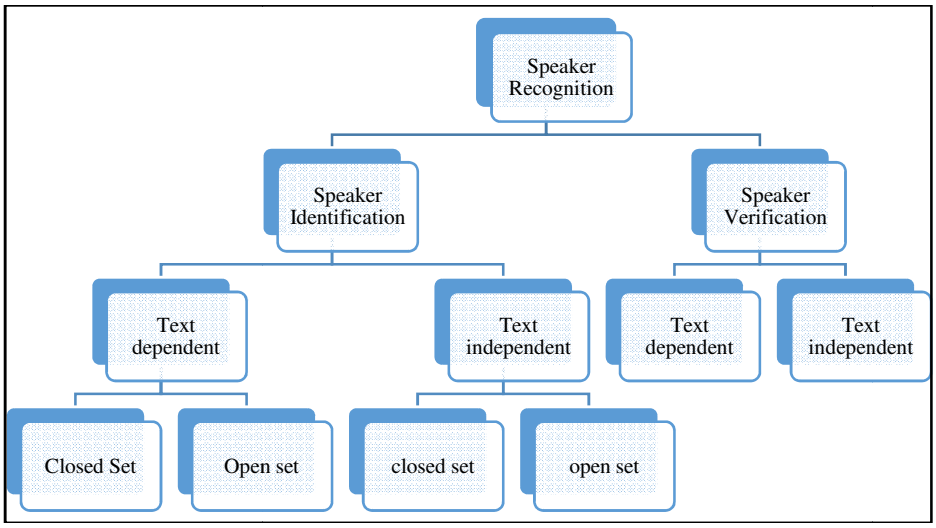


*Figure 1: Speaker Recognition*

**2. Tasks of a SRS**

A typical speaker recognition system involves several tasks and this section briefly explains each of them.

*2.1. Pre-processing*

The recorded speech signal may have some undesired noise due to additive acoustic noise or a reverberation or channel/media effects. Sometimes the environment may differ between the training and testing phase and all these may tends to decrease the performance of Speaker Recognition System. Removing these undesired noises and managing the mismatch in environment and channel is called pre-processing.

*2.2. Feature Extraction*

A speech signal carries information about the speech (what was spoken), speaker (who was spoken), language (the language that was spoken like English or Tamil etc.) and the emotion (happy or angry etc.). A Speaker Recognition System extracts the features that are uniquely identifies/verifies a speaker. The speech signal is converted into digital form and the feature extraction module extracts features that unique to recognize the speaker from the waveform. These speaker specific features can be broadly classified as "low-level" features and "high-level" features.

2.2.1. Low-level Features

Most of the speaker recognition system uses the spectral or the acoustic features of speech signals. These acoustic features represent the physical structure of the vocal tract and differ from person to person. It is also called as low-level features and preferred due to their low-complexity. The low-level features can be determined effectively from a very short overlapping frame of the speech signals, generally less than 30ms. The speech signals are quasi-stationary in nature but in a short duration (between 20ms and 30ms) its characteristics found to be stationary. Mel Frequency Cepstral Coefficient (MFCC), Linear Prediction Cepstral Coefficient (LPCC), and Perceptual Linear Prediction (PLP) are few examples of low-level features. These features are good for a system that has data come from the same training and testing environment or if there is no mismatch between the training and testing phase.

Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is the widely used low-level feature in speaker recognition systems. Human ears have varying bandwidth with frequencies and filters are spaced linearly at low frequencies and logarithmically at high frequencies. The block diagram indicating the steps involved in extracting the MFCC is given below:

The speech signal in divided into small frames of size 25ms to 40ms and frames are formed for every 10ms. The frame size is chosen in such a way that characteristics of speech signals do not vary much. A suitable windowing function, generally Hamming or Hanning window is performed on each frame to reduce the artifacts caused by the window size. A Fast Fourier transform is applied on the speech samples to calculate the magnitude spectrum and then the samples are processed by a bank of band-pass filters. The frequency bin is then transformed into mel-scale bins. This mel-frequency weighted magnitude spectrum is then processed by a non-linearity function and finally by a DCT. Ignoring few first and last DCT coefficients, remaining coefficients represent the MFCC features.

Linear Predictive Cepstral Coefficients (LPCC)

In LPCC method, the given speech sample can be approximated with the past 'n' speech samples. After framing and windowing the speech signal, either the autocorrelation or covariance methods will give this approximation.After segmenting the speech data into frames windowed frame is auto correlated to give

$$r_i(m) = \sum_{n=0}^{M-1-m} \overline{x}(n)\,\overline{x}(n+m)$$

2.2.2. High-level Features

Acoustic features of speech differ amongst individuals. These acoustic features include both learned behavioral features (e.g. pitch, accent) and anatomy (e.g. shape of the vocal tract and mouth) [10]. The properties of speech signal like pitch, tone, volume are unique to the individuals because these properties are depend on the size and shape of the mouth, vocal and nasal tract along with the size, shape and tension of the vocal tract. The pattern of words used, phone duration, stress pattern, intonation, idiolect, vocabulary, pitch, rate and rhythm of speech, formant frequencies, energy distribution of a longer frame are considered as "high-level" features. These features are captured from speech signals with time-scale more than few seconds and the computation complexity of these features are higher than that of low-level features.

A few researches have been done by combining the high-level features with low-level features using fusion techniques.

*2.3. Speaker Modeling*

The features extracted from the speech signals are represented in a feature vectors and these feature vectors are used to create a speaker's model. The number of reference templates need for efficient speaker recognition system depends upon the kind of features extracted. The speaker models can be broadly classified as generative or discriminative model.

2.3.1. Gaussian Mixture Model (GMM)

Assuming that feature vectors follow a Gaussian distribution, a Gaussian model characterizes mean and a deviation about the mean of a feature vectors. Given a feature vector xt, a GMM for a speaker is a weighted sum of N component densities, given as,

$p(x_t, \lambda) = \sum g_i \, N(x_t, \mu_i, \sum i)$. The parameters of a GMM are the mean vectors, covariance matrices, and mixture weights from all N component densities. If the number of mixtures is sufficiently enough (say 64 or more), the component densities can represent the individual speaker's broad phonetic class distributions.

The parameters of a GMM are estimated using the *expectation-maximization* (EM) algorithm [17][18]. It was observed that the EM algorithm guarantees monotonic convergence to the set of optimal parameters in only a few iterations. In the field of speaker recognition, GMM is the widely used model due to its better results in recognizing speakers. But the results of GMM can be relied upon only if we have enough training data to properly estimate the model parameters. Another drawback with GMM is that the mismatch in dataset between the training and testing session degrades the overall system performance. These problems can be solved by having huge volume of varied data, but in real-time it is difficult to get this.

The above mentioned drawbacks or limitations of GMM were overcome by an approach called Universal Background Model (UBM), which models all speakers other than the claimed speaker. Since UBM is trained by all of the feature space of all speaker data, the insufficient and unseen data problems were solved.

2.3.2. Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) is a statistical model in which the temporal productions are represented as first-order Markov process. Figure 2 shows a sample HMM which comprises of a sequence of states with a GMM associated with each state. Each state in the GMM represents a The stationary unit of the speech signal, called "tri-phone" is represented a particular state in the GMM. During the training phase HMM uses the Estimation-Maximization (EM) algorithm. In the evaluation phase, the most probable sequence of states/phones is estimated for a given speech signal. The scores calculated are accumulated to obtain the utterance and speaker specific likelihood. The HMMs are mainly used in text-dependent speaker verification systems since it depends on the phonic content of the speech signal.

2.3.3. Support Vector Machine (SVM)

The Support Vector Machine (SVM) derives an optimal separation between the target and non-target speakers by fitting a hyperplane. This hyperplane is chosen according to a maximum margin criterion [19]. According to MM criterion, a hyperplane that maximizes the Euclidean distance to the nearest data point on each side of the plane is chosen. One of the limitations of SVM is that, it is suitable for static data vectors. But the feature extractor, for a given speech utterance extracts a sequence of feature vectors instead of a single data vector. But this sequence of feature vectors can be converted, using a polynomial classifier [20], sequence kernel functions [21] or GMM super vectors [22], to a single data vector.

2.3.4. Vector Quantization (VQ)

A Vector Quantization (VQ) is a technique in which vectors of a large space is mapped to a finite number of regions in that space, called cluster. A cluster is represented by its centroid [9]. A group of all such centroids forms a codebook. Even though the codebook is smaller than the original sample, it still accurately represents a person's voice characteristics. Since the amount of data is significantly less, it reduces the amount of computations needed for comparison in later stages.

*2.4. Pattern Matching*

Pattern matching is the process of comparing a test speech signal against the previously stored template and determining the similarity score. Acceptance or rejection of the claim is then based upon whether the score exceeds the given threshold.

**3. Overview of basic Speaker Recognition Systems**

Plenty of research or activity has been done on speaker recognition and still many research works is going on improving the speaker recognition model with a lesser error rate or finding a robust speaker recognition model. This section discusses the conventional SR models and advancement in speaker recognition.

The cepstrum feature and pattern matching technique was first used in a text-dependent ASR system and the system resulted in 2% error rate for speaker verification and speaker identification. Even though the error rate was less, the system was tested with the population of size 10, and hence one could not rely on the effectiveness of the system [55]. Another SR system used Filter-bank features and DTW for the recognition model. This system has shown a better result of 0.8% error rate at 6s for the population size 200 [56]. Text-dependent SR system that included the cepstrum and LP features with autocorrelation and projected a long term statistics. The EER found in this system was 1% at 3s[57].

Another approach proposed in [type ref. number here] used Mel-cepstrum feature and the stochastic model HMM (GMM) for the recognition system. The system showed a very good result of 0.12% error rate @ 10s for the population of 138 people [58]. Mohd Zaizu Ilyas et al [59] presented a speaker verification system using a combination of Vector Quantization (VQ) and Hidden Markov Model (HMM) to improve the HMM performance. A Malay spoken digit database which contains 100 speakers was used for the testing and validation modules. It was shown that, by using that combination technique, a total success rate (TSR) of 99.97% was achieved and it was an improvement of 11.24% in performance compared to HMM. For speaker verification, true speaker rejection rate, impostor acceptance rate and equal error rate (EER) were also improved significantly compared to HMM.

| Method | FRR | FAR | TSR | ERR |
|---|---|---|---|---|
| **HMM** | 25.30% | 9.99% | 89.87% | 16.66% |
| **VQ + HMM** | 0.06% | 0.03% | 99.97% | 11.72% |

*Table 1*

This system has shown some improvement on a noise-free environment, nothing has been said about noisy environment.

In [60] the author proposed a text-independent speaker identification system based on Mel-Frequency Cepstrum Coefficient (MFCC) feature vectors and Hidden Markov Model (HMM) classifier. The implementation of the HMM was divided into two steps: feature extraction and recognition. In the feature extraction step, the paper reviews MFCCs by which the spectral features of speech signal can be estimated and how these features can be computed. In the recognition step, the theory and implementation of HMM were reviewed and followed by an explanation of how HMM can be trained to generate the model parameters using Forward-Backward algorithm and tested using forward algorithm. The HMM was evaluated using data of 40 speakers extracted from Switchboard corpus. Experimental results had shown an identification rate of about 84%.

When compared to the state of the SRS in seventies and eighties, today a more advanced and robust SRS are available even that can be used commercially

One of the promising developments in speaker recognition system was the introduction of *supervectors*. A supervector represents the utterances as a single vector generally by combining many smaller dimensional vectors into a single higher-dimensional vector. Several authors have studied different robust methods for both SVM and GMM models. The generalized linear discriminant sequence (GLDS) [36] is a simple SVM based supervector. This approach creates a supervector by explicitly mapping into kernel feature space using a polynomial expansion. The polynomial expansion includes either second- or third-order monomials before the dimensionality gets infeasible and hence the main drawback of GLDS method is it is difficult to control the dimensionality of the supervectors. A different approach was proposed by the same author in [37] [38] based on GMM, Gaussian supervector (GSV) kernel. The GSV kernel is derived by bounding the Kullback-Leibler (KL) divergence measure between the GMMs. Some researchers used Bhattacharya distance instead of KL [39]. In [40], MLLR is given as input to SVM and in [41] [42] high-level features are used in SVM.

**4. Issues or Challenges in SRS**

The several SRS that includes MFCC, LPCC, PLP features or speaker models like HMM, GMM, SVM, GMM-UBM, GMM-SVM produce high accuracy in clean conditions. But in real-time the results of SRS are often influenced by several factors like health or aging of a speaker [23], sleepiness [24], emotional state [25][26][27], phonetic variation [28], background noise [29][30][31] and transmission channel [32][33][34][35]. These factors can be put in under the two major classifications: *intra-speaker variability* and *inter-speaker variability*. In an inter-speaker variability, the efficiency of a Speaker Recognition System considerably reduces when the testing environment is different from training environment. This mismatch could be due to several factors like background noise, channel/handset distortion, room reverberation etc.,

**5. Robust Speaker Recognition**

The real-time speaker recognition system should address the various distortion factors there by increasing the robustness of the system. Approaches that have been developed to provide robustness against these effects can be broadly categorized into (i) Feature based compensation, (ii) Score based compensation and (iii) Model based compensation. Feature based compensation methods modify the features of noisy signal and then model a speaker. Score based methods remove the model score biases and shifts due to the mismatches. Model based compensation methods modify the trained models to learn the characteristics of noise and thus makes the decision making process more robust. The remaining part of this section explores the several robust speaker recognition systems and discusses their performances.

*5.1. Channel Robust*

Channel variability in one among the difficult challenges that SRS face. The most commonly used technique is Cepstral Mean Normalization (CMN).

A new cross-channel compensation technique was introduced for GMM-UBM systems in [3]. It includes wideband noise reduction, echo cancellation, a simplified feature-domain Latent Factor Analysis (LFA) and data-driven score normalization. They also developed a novel dynamic Gaussian selection algorithm to reduce the feature compensation time by more than 60% without any performance loss. By combining the above techniques, they were able to reduce the relative EER by 46% in cross-channel condition.

A different approach was used called Nuisance Attribute Projection (NAP) in [2]. In this approach few dimensions that are not needed to the SR system were removed and the results were promising.

To compensate the channel mismatch, most recent researches focus on the state-of-art technique called i-vectors which was initially proposed by Dehak et al in [33]. This approach combines both speaker information and channel differences into a single subspace from which the Baum-Welch first-order statistics are derived. From this high-dimensional supervector, a low dimensional fixed-length vector is estimated using MAP estimation. This low dimensional vector is called as I-vector. These I-vectors are then normalized by their mean and length. Probabilistic Linear Discriminant Analysis (PLDA), a generative factor approach is used to model the I-vectors. A log-likelihood ratio (LLR) is then computed between same versus different speaker hypothesis.

A variation to the PLDA model includes the channel estimate in the PLDA model for each test segment and as a result, shifts the scoring function to better match the testing channel [53]. The i-vector/PLDA based systems yields very good result for seen channel conditions, and for an unseen channel conditions, their performance degrades [54]. To minimize this unseen channel mismatch, Zhu et al., have proposed nearest neighbor based i-vector mean normalization (NN-IMN) and i-vector smoothing (IS) [54]. This approach can handle multiple unseen channels without explicit clustering or retraining. They also have shown that one can recover 46% of the total performance degradation with NN-based i-vectors.

The above discussed state-of-the-art methods for robust speaker recognition systems are mostly to compensate channel mismatches but not for additive background noise.

*5.2. Noise Robust*

Two model-based approaches called Spectral Equalization (SE) [43] and Spectral Subtraction (SS) [44] were proposed for compensating noise. The spectral subtraction method determines the clean speech spectra by subtracting the mean noise spectra from that of noisy speech spectra. The noise is assumed to be stationary in this method. The minimum mean squared error (MMSE) overcomes this limitation by correlating the frequency components [45] [46].

A simple and straight forward approach was proposed in [13] for an automated speaker recognition system under noisy environment. In this approach, wiener filter was used to remove background noise from the original speech utterances which was previously used in [14, 15, 16]. Speech end points detection and silence part removal algorithm has been used to detect the presence of speech and to remove pulse and silences in a background noise. Then the speech signal was segmented into overlapping frames and windowing techniques were applied. Features were extracted and then fed to the Discrete Hidden Markov Model for learning and classification.

In early 1990s, stereo-based (data that consists of simultaneous recordings of both the clean and noisy speech) compensation techniques were introduced in [47]. The Codeword Dependent Cepstral Normalization (CDCN) and Fast CDCN (FCDCN) [47] methods derive the Cepstral compensation vectors from stereo database and these vectors were applied to the training data to adapt to environmental changes. Fast codeword-dependent cepstral normalization (FCDCN) [48] was developed to provide a form of compensation that provides greater recognition accuracy than CDCN but in a more computationally-efficient fashion than the CDCN algorithm. The alignment-based codeword dependent cepstral normalization algorithm (ACDCN) [49] which aims to alleviate the acoustical mismatch that occurs when the speech recognizer faces environmental conditions not observed in the training data. ACDCN is based on the linear channel model of the environment originally proposed by Acero and on the CDCN solution to this model. The drawback of these algorithms was lack of learning the variance of the distribution. As an alternate, a new family of environmental compensation algorithms called Multivariate Gaussian-based Cepstral Normalization (RATZ) [50]. RATZ assumes that the effects of unknown noise and filtering on speech features can be compensated by corrections to the mean and variance of components of Gaussian mixtures and an efficient procedure for estimating the correction factors is provided. The RATZ algorithm was implemented to work with or without the use of "stereo" development data that had been simultaneously recorded in the training and testing environments. "Blind" RATZ partially overcomes the loss of information that would have been provided by stereo training through the use of a more accurate description of how noisy environments affect clean speech. These algorithms construct rigid functions based on approximations to known causes of distortion, such as additive noise and linear convolutional channels. The Stereo-based Piecewise Linear CompEnsation for Environments (SPLICE) is a framework used to model and remove the effect of any consistent degradation of speech cepstra. It learns a joint probability distribution of noisy and clean cepstra, and uses this distribution to infer clean speech estimates from noisy inputs. However SPLICE does not include any assumptions about how noisy cepstra are produced from clean cepstra, and can model any combination of these affects as well as others, including nonlinear and possibly non-stationary distortions. SPLICE approach gave a better result in robust speech recognition when compared to the previous algorithms. In [51], multi-environment models based linear normalization (MEMLIN) was proposed based on MMSE estimation. This algorithm learns the difference between clean and noisy feature vectors associated to a pair of Gaussians (one for a clean model, and the other for a noisy model), for each basic environment. This knowledge, the associated Gaussians, the conditional probability between clean and noisy Gaussians, and the environment are the data used to compensate the mismatch between clean and noisy vectors. Another development in stereo-based algorithm that has received attention of researchers is Stereo-based Stochastic Mapping (SSM) [52]. Both

the clean and noisy channels are used to form a large augmented space and the statistical models are built in this new space. In the testing phase, both the observed noisy features and the joint statistical model are used to predict the clean observations.

One of the recent approaches towards robust speaker recognition is missing data method. The missing data approach compensates against arbitrary disturbances within a speech signal, and is thus capable of dealing with the problem of environmental noise [4]. The approach is based on a time-frequency analysis of the input speech signal, and the subsequent quantification of noise in each individual time-frequency point.

These approaches are critically dependent on the accuracy with which individual time-frequency regions within a speech signal can be identified as speech or noise dominant. In practical situations the absence of a priori noise knowledge requires an estimation of these reliability decisions in the form of a reliability mask. In the past approaches the reliability mask estimation for speaker recognition were done using SNR-based methods, auditory and perceptual criteria and classification-based techniques. While SNR-based methods offer simplicity by attempting a direct estimation of the noise spectra from speech free regions, their performance is generally inferior to auditory and classifier techniques, which identify regions of speech dominance by utilizing perceptual cues (such as locality or pitch information) and specifically designed features respectively. Regardless of the technique used to estimate the reliability decisions, in difficult noise conditions these bottom-up estimation methods will produce reliability mask errors which adversely affect recognition performance. So the authors examined the top-down approaches as a solution to address the vulnerability of traditional missing data systems to mask errors.

A similar approach has been studied in [5], finding a noise-robust speaker recognition combining missing data and UBM. Their experiments showed that the usage of a UBM in combination with missing data recognition yields substantial improvements in recognition performance, especially in the presence of highly non-stationary background noise at low SNRs.

To overcome the limitations of noise compensation techniques, a Missing Frequency (MF) [10] approach was used. The MF approach can handle any unknown noise and does not require a priori knowledge of noise that distorts the speech signal. This approach assumes noise affects the time-frequency (t-f) regions of the speech spectrum in different ways, detects the spectrum corruption level and chooses a segment of spectrum that is reliable enough to be used in recognition. The MF approach provides robustness in SR in noisy environment but its efficiency depends upon the mask estimation accuracy. The Signal to noise ratio (SNR) is a measure to quantify how much a signal is corrupted by noise. Most of the SRS uses SNR to estimate the mask.

A variation in the MF approach called "Features classification (FC) criterion" [11], improves over the SNR criterion making use of several other complimentary features. The MF-FC approach used several spectro-temporal measures that are independent of one another and hence even if feature is corrupted by noise, still the other features could ensure that reliability decision remains consistent. However, the computational cost is higher than the approach that using SNR.

Additive noise reduction methods usually have a tradeoff between the amount of noise reduction and speech distortion induced due to processing of a speech signal. M.J. Alam et al proposed robust compressive gammachirp filterbank cepstral coefficient (RCGCC) [12] feature extractor for robust speech recognition. The RCGCC feature extractor, usual preprocessing steps were applied, then short-time Fourier Transform (STFT) analysis is performed using a finite duration (25ms) Hamming window with a frame shift of 10ms to estimate the power spectrum of the signal. Compressive gammachirp filter-bank (cGCFB) integration is performed on both speech and noise power spectra for auditory spectral analysis. A sigmoid-shaped weighting rule is applied to enhance the auditory spectrum. The 13-dimensional static features, obtained after applying power function nonlinearity with a coefficient of 1/15 and the discrete cosine transform (DCT), are normalized using the short-time cepstral mean and scale normalization (STCMSN) technique.

**6. Conclusion**

Most of the automatic speaker recognition system yields good performance results, in a clean environment. Even though, lot of research have been done on to make the speaker recognition system robust, still it is a challenging task to develop a better pattern matching algorithm for a robust SRS. In this review article, we have discussed some of approaches developed by the research community in building a noise and channel robust systems. However the current biggest challenge is how to adapt the system to a totally new and unseen condition. The other related challenges are the production of reliable recognition results using the high level features such as pronunciation, speaking style, quality of voice etc., and another big challenge is developing an ASRS in a mobile environment. Lot of research activity is going on to improve the efficiency of SRS in a mobile environment that is tested with several mobile speech data.

We can implement a new SRS or develop a better pattern matching algorithm in the lab, however it is highly difficult to use it in a really challenging environment. But only when a SRS is tested in such conditions, it will become usable in our day-to-day life.

**7. References**


i. Joseph P. Campbell, "Speaker Recognition: A Tutorial", Proceedings of the IEEE, VOL. 85, NO. 9, SEPTEMBER 1997.
ii. Alex Solomonoff, W.M. Campbell, and Ian Boardman, "Advances in Channel Compensation for SVM speaker recognition", in Proc., ICASSP 2005, pp I-629-632
iii. Yuxiang Shan and Jia Liu, "Robust speaker recognition in cross-channel condition based on Gaussian mixture model" in Multimed Tools Appl 2011Springer, 52:159-173
iv. Roberto Togneri and Daniel Pullella, "An Overview of Speaker Identification: Accuracy and Robustness Issues", IEEE Circuits and Systems Magazine Second Quarter 2011, pp 23-60

v. 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, pp 108-121
vi. Taufiq Hasan and John H.L. Hansen, "A Study on Universal Background Model Training in Speaker Verification", IEEE Transactions on Audio, Speech, and Language Processing, Vol 19, No 7, 2011, pp 1890-1899
vii. Xianyu Zhao and Yuan Dong, "Variational Bayesian Joint Factor Analysis Models for Speaker Verification", IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No 3, 2012, pp 1032-1042
viii. Mitchell McLaren and David van Leeuwen, "Source-Normalized LDA for Robust Speaker Recognition using i-Vectors From Multiple Speech Sources", IEEE Transactions on Audio, Speech, and Language Processing, Vol 20, No 3, 2012, pp 755-766
ix. F.K. Soong, A.E. Rosenberg, L.R. Rabiner and B.H. Juang, "A Vector Quantization approach to Speaker Recognition", Floria: ICASSP Vol.1, 1985, pp. 387-390
x. Raj, B., Stern, R., "Missing-feature approaches in speech recognition", IEEE Signal Processing Magazine, 2005
xi. Dayana Ribas Gonzalez, Jose Ramon Calvo de Lara, "Feature classification criterion for missing features mask estimation in robust speaker recognition", SIViP, Vol 8, 2012 pp 365-375
xii. M J Alam, P Kenny, D O'Shaughnessy, "Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique", Digital Signal Processing 29, Elsevier, 2014, pp 147-157
xiii. Md. Rabiul Islam, Md. Fayzur Rahman, "Text dependent Speaker Identification System using Discrete HMM in Noise", IJCA, Vol 21, No 3, 2011
xiv. Simon Doclo, Marc Moonen, "On the Output SNR of the Speech-Distortion Weighted Multichannel Wiener Filter", IEEE Signal Processing Letters, Vol. 12, 2005
xv. Wiener N., "Extrapolation, Interpolation an Smoothing of Stationary Time Series with Engineering Applications", Wiely, Newyork, 1949
xvi. Wiener N., Paley, R.E.A.C., "Fourier Transforms in the Complex Domains", American Mathematical Society, Providence, RI, 1934
xvii. J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Int. Comput. Sci. Inst. [Online]. 4. Available: http://ssli.ee.washington.edu/people/bilmes/mypapers/em.pdf, 1998
xviii. D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process., vol. 3, no. 1, pp. 72–83, 1995
xix. C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining Knowl. Discov., vol. 2, no. 2, pp. 121–167, 1998.
xx. V. Wan and W. Campbell, "Support vector machines for speaker verification and identification," in Proc. IEEE Neural Networks Signal Process. Workshop, 2000, vol. 2, pp. 775–784.
xxi. W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," Comput. Speech Lang., vol. 20, no. 2–3, pp. 210–229, 2006.
xxii. W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters vol. 13, no. 5, pp. 308–311, 2006.
xxiii. F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification in score ageing- quality classification space," Comput. Speech Lang., vol. 27, no. 5, pp. 1068–1084, Aug. 2013
xxiv. T. Rahman, S.Mariooryad, S. Keshavamurthy, G. Liu, J. H. L. Hansen, and C. Busso, "Detecting sleepiness by fusing classifiers trained with novel acoustic features," in Proc. Interspeech, Florence, Italy, Aug. 2011, pp. 3285–3288.
xxv. B. Schuller, S. Steidl, A. Batliner, E. Noth, A. Vinciarelli, F. Burkhardt, R. V. Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The interspeech 2012speaker trait challenge," in Proc. Interspeech, Portland, OR, USA, Sep. 2012, pp. 254–257.
xxvi. G. Liu, Y. Lei, and J.H. L. Hansen, "A novel feature extraction strategy for multi-stream robust emotion identification," in Proc. Interspeech, Makuhari, Japan, Sep. 2010, pp. 26–30.
xxvii. J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," Speech Commun., vol. 20, no. 1, pp. 151–173, Nov. 1996.
xxviii. R. J. Vogt, B. J. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in Proc. Interspeech, Brisbane, Australia, Sep. 2008, pp. 853–856.
xxix. G. Liu, Y. Lei, and J. H. L. Hansen, "Robust feature front-end for speaker identification," in Proc. ICASSP, Kyoto, Japan,Mar. 2012, pp. 4233–4236.
xxx. Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in Proc. ICASSP, Vancouver, BC, Canada, May 2013, pp. 6788–6791.
xxxi. Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in Proc. ICASSP, Kyoto, Japan, Mar. 2012, pp. 4253–4256.
xxxii. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 4, pp. 1448–1460, May, May 2007.
xxxiii. M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-Vector extractor suitable for speaker recognition with both microphone and telephone speech," in Proc. Odyssey, Brno, Czech Republic, Jun. 2010, pp. 28–30.

xxxiv. N. Dehak, Z. N. Karam, D. A. Reynolds, R. Dehak, W. M. Campbell, and J. R. Glass, "A channel-blind system for speaker verification," in Proc. ICASSP, Prague, Czech Republic, 2011, pp. 4536–4539.

xxxv. L. Burget, N. Brummer, and D. Reynolds, "Robust speaker recognition over varying channels," in Johns Hopkins University CLSP Summer Workshop Rep., 2008 [Online]. Available: www.clsp.jhu.edu/workshops/ws08/documents/jhu_report_main.pdf

xxxvi. Campbell, W., Campbell, J., Reynolds, D., Singer, E., Torres-Carrasquillo,P., 2006a. Support vector machines for speaker and language recognition. Computer Speech Language 20 (2–3), 2006, pp. 210–229.

xxxvii. Dehak, N., Chollet, G., "Support vector GMMs for speaker verification" In: Proc. IEEE Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2006), San Juan, Puerto Rico, June 2006.

xxxviii. Lee, K.-A., You, C., Li, H., Kinnunen, T., "A GMM-based probabilistic sequence kernel for speaker verification" In: Proc. Interspeech 2007 (ICSLP), Antwerp, Belgium, August 2007, pp. 294 – 297.

xxxix. You, C., Lee, K., Li, H., "An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition", IEEE Signal Process. Letter. 16 (1), 2009, pp. 49–52.

xl. Stolcke, A., Kajarekar, S., Ferrer, L., Shriberg, E., "Speaker recognition with session variability normalization based on MLLR adaptation transforms" IEEE Trans. Audio, Speech Language Process. 15 (7), 2007.

xli. Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A., "Modeling prosodic feature sequences for speaker recognition" Speech Comm. 46 (3–4), 2005, pp. 455–472.

xlii. Campbell, W., Campbell, J., Reynolds, D., Jones, D., Leek, T., "Phonetic speaker recognition with support vector machines" In: Thrun, S., Saul, L., Schokopf, B. (Eds.), In: Advances in Neural Information Processing Systems, Vol. 16. MIT Press, Cambridge, MA, 2004.

xliii. T.G. Stockham, T.M. Cannon, R.B. Ingebretsen, "Blind deconvolution through digital signal processing" Proc. IEEE 63(4), 1975, pp 678–692

xliv. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoust. Speech Signal Process. 27(2), 1979, pp113–120

xlv. A. Erell, M.Weintraub, "Spectral estimation for noise robust speech recognition", Proceedings of DARPA Speech and Natural Language Workshop, Philadelphia, 1989.

xlvi. Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator" IEEE Trans. Acoust. Speech Signal Process. 33(2), 1985, pp 443–445.

xlvii. A. Acero, "Acoustical and environmental robustness in automatic speech recognition", PhD thesis, Carnegie Mellon University, 1990.

xlviii. A. Acero, R.M. Stern, Environmental robustness in automatic speech recognition, in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '90), Albuquerque, 1990, vol. 2, pp. 849–852

xlix. Huerta, J.M., "Alignment-based codeword-dependent cepstral normalization," Speech and Audio Processing, IEEE Transactions on , vol.10, no.7, 2002, pp.451-459.

l. Moreno, P.J.; Raj, B.; Gouvea, E.; Stern, R.M., "Multivariate-Gaussian-based cepstral normalization for robust speech recognition," Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on , vol.1, no., pp.137,140 vol.1, 1995

li. Buera, L.; Lleida, E.; Miguel, A.; Ortega, A., "Multi-environment models based linear normalization for speech recognition in car conditions," Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on , vol.1, no., pp.I,1013-16 vol.1, 2004.

lii. M. Afify, X. Cui, Y. Gao, Stereo-based stochastic mapping for robust speech recognition. IEEE Trans. Audio Speech Lang. Process. 17(7), 2009, pp 1325–1334

liii. Liping Chen, Kong Aik Lee, Bin Ma, Wu Guo, Haizhou Li, and Li Rong Dai, "Channel Adaptation of PLDA for text-independent speaker verification", ICASSP 2015

liv. Weizhong Zhu, Sadjadi, Jason W. Pelecanos, "Nearest Neighbor Based i-vector Normalization for Robust speaker recognition under unseen channel conditions", ICASSP, 2015

lv. B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Amer., vol. 55, no. 6, pp. 1304–1312, 1974.

lvi. G. R. Doddington, "Speaker recognition—Identifying people by their voices," Proc. IEEE, vol. 73, pp. 1651–1664, Nov. 1985.

lvii. J. Attili, M. Savic, and J. Campbell, "A TMS32020-based real time, text-independent, automatic speaker verification system," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp. 599–602, New York, 1988.

lviii. D. Reynolds and B. Carlson, "Text-dependent speaker verification using decoupled and integrated speaker and speech recognizers," in Proc. EUROSPEECH, pp. 647–650,Madrid, Spain, 1995.

lix. Mohd Zaizu Ilyas, Salina Abdul Samad, Aini Hussain and Khairul Anuar Ishak, "Speaker Verification using Vector Quantization and Hidden Markov Model", The 5th Student Conference on Research and Development – SCOReD, Malaysia, 2007.

lx. Sayed Jaafer Abdallah, Izzeldin Mohamed Osman and Mohamed Elhafiz Mustafa, "Text-Independent Speaker Identification Using Hidden Markov Model", World of Computer Science and Information Technology Journal (WCSIT), Vol. 2, No. 6, 203-208, 2012.