Speaker Recognition: A Survey

Prof. Vani A. Hiremani
Pune Institute of Computer Technology, Pune

Abstract- Today many people have access to their company’s information system by logging in from home. Also Internet services and telephone banking are widely used by corporate and private sectors. Therefore to protect one's resources or information with simple password is not reliable and secure in the world of today. Biometric systems are used for recognizing a user based upon his/her unique physiological and/or behavioral characteristics. Voice signal as a unique behavioral characteristic is presented for speaker verification over telephone lines. Voice recognition system is capable of verifying a speaker among the group of speakers.

Keywords— Speaker recognition types, features extraction pros and cons.

I. INTRODUCTION

Speaker recognition technology is the most potential technology to create new services that will make our everyday lives more secured. Another important application of speaker recognition technology is for forensic purposes. Speaker recognition has been an appealing research field for the last decades which still yields a number of unsolved problems.

Speaker recognition is basically divided into two-classification: speaker recognition and speaker identification and it is the method of automatically identify who is speaking on the basis of individual information integrated in speech waves. Speaker recognition is widely applicable in use of user's voice to verify their identity and control access to services such as banking by telephone, database access services, voice dialing telephone shopping, information services, voice mail, security control for secret information areas, and remote access to computer. The main aim of this work is surveying whole process of speaker identification, which consists of comparing a speech signal from an unknown speaker to a database of known speaker. The system can recognize the speaker, which has been trained with a number of speakers. The speaker identification is the process of determining which registered speaker provides a given speech. On the other hand, speaker verification is the process of rejecting or accepting the identity claim of a speaker.

In most of the applications, voice is use as the key to confirm the identities of a speaker is classified as speaker verification. Speaker recognition can also divide into two methods, text- dependent and text independent methods. In text dependent method, the speaker may say key words or sentences those having the same text for both training and recognition trials. Whereas in the text independent does not rely on a specific text being speaks. Formerly text dependent methods were widely in application, but later text independent is in use. Both text dependent and text independent methods share a problem however.

By playing back the recorded voice of registered speakers this system can be easily deceived. There are different technique is used to cope up with such problems. Such as a small set of words or digits are used as input and each user is provoked to thorough a specified sequence of key words that is randomly selected every time the system is used.

Speaker Recognition

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves, which can be classified into identification and verification. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers.

In this way, speaker recognition technology is expected to create new services that will make our daily lives more convenient. Another important application of speaker recognition technology is for forensic purposes. Figure.1 shows the basic structures of speaker identification and verification systems.

Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Most applications in which a voice is used as the key to confirm the identity of a speaker are classified as speaker verification.
Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to say key words or sentences having the same text for both training and recognition trials, whereas the latter do not rely on a specific text being spoken. Both text-dependent and independent methods share a problem however. These systems can be easily deceived because someone who plays back the recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered speaker. To cope with this problem, there are methods in which a small set of words, such as digits, are used as key words and each user is prompted to utter a given sequence of key words that is randomly chosen every time the system is used. Yet even this method is not completely reliable, since it can be deceived with advanced electronic recording equipment that can reproduce key words in a requested order. Therefore, a text-prompted (machine-driven-text-dependent) speaker recognition method has recently been proposed.

Text-Dependent Speaker Recognition Methods

Text-dependent methods are usually based on template-matching techniques. In this approach, the input utterance is represented by a sequence of feature vectors, generally short-term spectral feature vectors.

Text-Independent Speaker Recognition Methods

One of the most successful text-independent recognition methods is based on vector quantization (VQ). In this method, VQ codebooks consisting of a small number of representative feature vectors are used as an efficient means of characterizing speaker-specific features. A speaker-specific codebook is generated by clustering the training feature vectors of each speaker. In the recognition stage, an input utterance is vector-quantized using the codebook of each reference speaker and the VQ distortion accumulated over the entire input utterance is used to make the recognition decision.

Text-Prompted Speaker Recognition Methods

In the text-prompted speaker recognition method, the recognition system prompts each user with a new key sentence every time the system is used and accepts the input utterance only when it decides that it was the registered speaker who repeated the prompted sentence. The sentence can be displayed as characters or spoken by a synthesized voice. Because the vocabulary is unlimited, prospective impostors cannot know in advance what sentence will be requested.

Literature survey

In recent years many of the articles were found which are dealing with the issues and challenges in speaker recognition. Some of the articles are surveyed and described below. In 2012, Conventional speaker recognition systems perform poorly under noisy conditions. Inspired by auditory perception, computational auditory scene analysis (CASA) [1] typically segregates speech by producing a binary time–frequency mask. Research investigated CASA for robust speaker identification. To deal with noisy speech, apply CASA separation and then either reconstruct or marginalize corrupted components indicated by a CASA mask. To tackle, problem with the systems that do not perform well under noisy conditions because extracted features are distorted by noise, causing mismatched likelihood calculation robustness problem, speech enhancement methods that are widely used in speech recognition, such as spectral subtraction, have been explored for robust speaker recognition.
The proposed system uses CASA as a front-end processor for robust SID. Figure 2 presents the overall system. Input speech is decomposed using a gamma tone filter bank and subsequent time windowing to generate a time sequence of GF’s, feed the input signal to a CASA system that computes a binary mask corresponding to the target speech. An important finding in this work is that GFCC features outperform conventional MFCC features under noisy conditions. MFCC is obtained by a discrete Fourier transform (DFT), followed by a conversion to the Mel-frequency scale with a bank of triangular filters. Applying DCT to the log energy of the filter output produces MFCC. There are two main differences between GFCC and MFCC. First, GFCC uses a gamma tone filter bank whereas MFCC uses a triangular filter bank applied to DFT. Gamma tone filters constitute a more accurate model of cochlear filtering than triangular filters. Second, a log operation is applied in deriving MFCC whereas a cubic root operation is used in GFCC derivation. In 2012, the use of Independent Component Analysis (ICA) and Principal Component Analysis (PCA) techniques [2] to reduce the dimensionality of high-level LVCSR features are proposed and at the same time to enable modeling those with state-of-the-art techniques like Probabilistic Linear Discriminant Analysis or Pairwise Support Vector Machines (PSVM). The high-level features are the coefficients from Constrained Maximum-Likelihood Linear Regression (CMLLR) and Maximum-Likelihood Linear Regression (MLLR) transforms estimated in an Automatic Speech Recognition (ASR) system. The results of this work show that MLLR features compressed with Probabilistic Principal Component Analysis (PPCA) and combined with generative cepstral models like PLDA can achieve better performance than NAP–SVM based techniques. MLLR and its variant CMLLR are techniques used to adapt speaker-independent models on the small amount of available speaker-specific data.

MLLR is a set of linear transforms, operating in the space of the model parameters, which maximizes the likelihood of the adaptation data by rotating all HMM model parameters. The transforms are estimated using the EM algorithm. Rank normalization: In SVM–NAP based systems rank normalization can be useful to increase performance of the system. For every dimension, a feature value is replaced by the position the feature would have occupied in the ordered set of values taken from a background training set. Probabilistic Principal Component Analysis: Principal Component Analysis is a well-known feature reduction technique. PCA can be defined as the linear projection that minimizes the average reconstruction cost for the data, where the cost is given by the mean squared distance between data points and their projections. The probabilistic model is given by,

$$x = Ws + \epsilon$$

Where $x$ is assumed to be a zero–mean $d$–dimensional observable random variable, $s$ is a $p$–dimensional Gaussian distributed latent random variable represents Gaussian distributed random noise. Independent Component Analysis: Independent Component Analysis is a technique that allows to linearly transforming a multidimensional random vector into statistically independent components. ICA can be interpreted as an extension to PCA, in the sense that PCA looks for dimensions which decorrelate data, while ICA looks for dimensions which make data independent. In 2012, the problem of speaker verification is addressed where additive noise is present in the enrollment and testing utterances [3]. Here it shows how the current state-of-the-art framework can be effectively used to mitigate this effect. First look at the degradation a standard speaker verification system is subjected to when presented with noisy speech waveforms. Research has designed and generated a corpus with noisy conditions, based on the NIST SRE 2008 and 2010 data, built using open-source tools and freely available noise samples. Then they shown how adding noisy training data in the current i-vector based approach followed by probabilistic linear discriminant analysis (PLDA) can bring significant gains in accuracy at various signal-to-noise ratio (SNR) levels. This work is focused on the robustness of speaker verification systems under noisy conditions, and how the proposed paradigm can help compensate for the observed degradation. A straightforward approach to achieving noise robustness in systems would be to add noisy data in all stages of the system: UBM, i-vector extractor, and LDA/PLDA.
UBM and i-vector extractor training are computationally expensive stages taking lots of memory and CPU resources. On the other hand, LDA/PLDA training is very fast by comparison. Initial attempts at adding noisy data to UBM and i-vector training have shown very small gains. Hence, in this work, only noisy training data added into the LDA/PLDA stage. The Weighted Linear Discriminant Analysis (WLDA) technique [4] is introduced in 2012, for the purposes of improving i-vector speaker verification in the presence of high inter-session variability. By taking advantage of the speaker discriminative information that is available in the distances between pairs of speakers clustered in the development i-vector space, the WLDA technique is shown to provide an improvement in speaker verification performance over traditional Linear Discriminant Analysis (LDA) approaches. Based upon the results presented within this work using the NIST 2008 Speaker Recognition Evaluation dataset, it is believed that both WLDA and WSNLDA are viable as replacement techniques to improve the performance of LDA and SNLDA-based i-vector speaker verification. In this work author has proposed to investigate a new LDA technique, based upon the weighted pair-wise Fisher criteria. This technique, known as Weighted LDA (WLDA), takes advantage of the discriminative information between pairs of classes, or speakers for our application, in the between-class scatter that has not yet been investigated for i-vector-based speaker verification. Finally, Euclidean distance weighted based WLDA performed better than Bayes Error weighted based WLDA. WLDA achieved 10% improvement over standard LDA under \textit{interview-interview} condition. However the marginal improvement of the performance in the telephone-telephone condition is likely to be due to the larger number of low session-count speaker recordings in the telephone development data, causing poorly estimated class means to reduce the quality of the estimations of the between-class scatter. In 2011, the main paradigms for speaker identification are coined and recent work on missing data methods to increase robustness [5]. The feature extraction, speaker modeling and system classification are discussed. Evaluations of speaker identification performance subject to environmental noise are presented. While performance is impressive in clean speech conditions, there is rapid degradation with mismatched additive noise. Missing data methods can compensate against arbitrary disturbances and remove environmental mismatches. In \textit{speaker identification} human speech from an individual is used to identify who that individual is. There are two distinct operational phases.

In \textit{training} (also called \textit{enrolment}) the speech from each known, verified speaker, for all speakers that need to be identified, is acquired to build (train) the model for that speaker. In \textit{testing} the true operation of the system is carried out where the speech from an unknown utterance is compared against each of the trained speaker models. In \textit{closed-set identification} the unknown individual belongs to a pre-existing pool or database of speakers (speaker models) and the problem then becomes that of choosing which speaker from the pool the unknown speech is derived from. The main performance measure of such systems is the identification rate (percentage of correct identification averaged across all speakers in the pool). In \textit{open-set identification} the unknown individual can come from the general population. The first task of an open-set identification system is to detect whether the speaker belongs to the pool or database of known speakers, if not, that speaker is rejected; otherwise, closed-set identification is carried out. In \textit{speaker verification} human speech from an individual is used to verify the claimed identity of that individual. The applicability of score and feature normalization techniques to parts-based Gaussian mixture model (GMM) [6] face authentication was proposed in 2011. In particular, research proposes to utilize techniques that are well established in state-of-the-art speaker authentication, and apply them to the face authentication task. For score normalization, T-, Z- and ZT-norm techniques are evaluated. For feature normalization, research proposes a generalization of feature warping to 2D images, which is applied to discrete cosine transform (DCT) features prior to modeling. This work focuses on robust techniques for normalization in two stages that have so far received less attention, that is, normalization of the features and normalization of the output scores. Gaussian Mixture Modeling For Face Authentication: The parts-based topology using Gaussian mixture modeling (GMM), method decomposes the face into an overlapping set of blocks, each of which is then considered to be a separate observation of the same signal (the face). This approach relies on estimating the distribution of features using a GMM for each subject, then performing authentication by calculating a likelihood ratio between the subject model and a universal background model (UBM). The rest of this section describes the main processing stages of the framework, including image registration, pre-processing, feature extraction and classification. Image Registration and Pre-processing: The image is converted to greyscale, cropped and registered using manually or automatically-localized eye positions.
In this work, experiments use manually-annotated eye positions. The resulting image is 64 _ 80 pixels with a distance between the eyes of 33 pixels, where the two eyes are aligned on the horizontal axis and the center of each eye is located 16 pixels down and 16 pixels in from the border of the image. Each cropped image is then processed using Tan & Triggs normalization, which consists of gamma correction, difference of Gaussian (DoG) filtering then contrast equalization. Feature Extraction: Feature extraction consists of segmenting a pre-processed image into a set of overlapping blocks and extracting a feature vector of 2D-Discrete Cosine Transform (2D-DCT) coefficients from each block. Blocks are exhaustively sampled from the image that is, sampled from the image with a step size of 1 pixel. The pixel values in each block are then mean and variance-normalized. From each of the K blocks in an image, we retain only the subset of D 2D-DCT coefficients that correspond to the low frequency range, since they are less susceptible to noise. Score Normalization: Score normalization aims to counteract statistical variations in output scores due to changes in the conditions across different enrolment and probe samples. This is achieved by scaling distributions of system output scores to better facilitate the application of a single, global threshold for authentication. Feature Normalization: Two of the most successful approaches to feature normalization in the field of speaker authentication are mean and variance normalization (MVN) and the more advanced technique of feature warping (FW). In future work, this research aim to apply feature warping both in the space and time dimensions to face videos. This could improve the robustness of visual features to noise that is time-varying, for example, variations in pose, illumination and expression throughout a video. It would also be interesting to see if the results in this work generalize to approaches other than GMM based systems. In 2011, a new low-dimensional speaker- and channel-dependent space is defined using a simple factor analysis [7]. This space is named the total variability space because it models both speaker and channel variability’s. Two speaker verification systems are proposed which use this new representation. First- system is a support vector machine-based system that uses the cosine kernel to estimate the similarity between the input data. Second- system directly uses the cosine similarity as the final decision score. This research tested three channel compensation techniques in the total variability space, which are within-class covariance normalization (WCCN), linear discriminate analysis (LDA), and nuisance attribute projection (NAP). And it is found that the best results are obtained when LDA is followed by WCCN. It is achieved an equal error rate (EER) of 1.12%.

This work proposed a new way of combining JFA(Joint Factor Analysis) and SVMs(Support Vector Machines) for speaker verification. It consists in directly using the speaker factors estimated with JFA as input to the SVM. It is tested several kernels and the best results were obtained using the cosine kernel when within-class covariance normalization (WCCN) is also used to compensate for residual channel effects in the speaker factor space. Front-End Factor Analysis: This section presents two new speaker verification systems which use factor analysis as a feature extractor. The first system is based on support vector machines and the second one uses the cosine distance value directly as a final decision score. This work presented a new speaker verification system where factor analysis is used to define a new low-dimensional space that models both speaker and channel variabilities. Two new scoring methods have been proposed based on the cosine kernel in the new space. The first approach uses a discriminative method, SVM, and the second one uses the cosine distance values directly as decision scores. The latter approach makes the decision process less complex because there is no speaker enrollment as opposed to the classical methods. The main difference between the classical use of joint factor analysis for speaker verification and approach is that research addressed the channel effects in this new low-dimensional i-vectors space rather than in the high-dimensional GMM mean super vector space. Research tested three different techniques to compensate for the intersession problem: linear discriminant analysis, nuisance attribute projection and within-class covariance normalization. A new method –named Transformation Network features with SVM modeling [8] is proposed– in order to become language independent and overcome the need for accurate speech recognition. This is accomplished by using a set of parallel acoustic models in several different languages to obtain a high-dimensional Parallel Transformation Network feature vector for speaker characterization.TN Features For Speaker Recognition ANN/HMM speech recognition and phone transcriptions. The core speech recognizer uses Multiple Layer Perception (MLP) networks that act as phoneme classifiers for estimating the posterior probabilities of a single state Markov chain mono-phone model. The baseline system for narrowband data combines four MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivatives), log-RelAtive SpecTrAl features (RASTA, 13 static + first derivatives), Modulation SpectroGram features (MSG, 28 static) and the advanced Font-End from ETSI features (ETSI, 13 static + first and second derivatives).
MLP/HMM Speaker Adaptation: The Transformation Network technique employs a trainable linear input network to map the speaker dependent (SD) input vectors to the characteristics of the speaker independent (SI) connectionist system. In order to train the TN for a new speaker, the weights of the mapping are initialized to an identity matrix. This guarantees that the SI model is the initial point prior to adaptation. During training, the output error of the posterior probabilities is calculated and back-propagated as usual in MLP training. But the SI part is kept frozen and weight adaptation is performed only in the new transformation network. TN feature extraction and SVM speaker modeling: Linear speaker dependent mappings for each data segment are independently trained and consequently a speaker adapted transformation matrix is obtained for each segment. In order to avoid capturing too much information of the background or channel conditions, long segments of silence were removed from the adaptation data. Thus, the dimensionality of the linear mapping is reduced to just [Nfeat, Nfeat]. The coefficients from the linear mapping obtained for each speaker are concatenated in a vector together with the segment mean and variance statistics of the feature data. Speaker verification is a popular biometric identification technique used for authenticating and monitoring human subjects using their speech signal. The method is attractive for two main reasons: (a) It does not require direct contact with the individual, thus avoiding the hurdle of “perceived invasiveness” inherent in many biometric systems like iris and finger print recognition systems; (b) It does not require deployment of specialized signal sensors as microphones are now ubiquitous on most portable devices (cellular phones, PDAs and laptops).Operation of a speaker verification system typically consists of two distinct phases: (a) An enrollment phase where parameters of a speaker specific statistical model are determined using annotated (pre-labeled) speech data; and (b) A verification phase where an unknown speech sample is authenticated using the trained speaker specific model. Both the phases are shown in Figure 3, where the speech signal is first sampled, digitized and filtered before a feature extraction algorithm computes salient acoustic features from the speech signal. The next step in the enrollment phase uses the extracted features to train a speaker specific statistical model [9].
However, the biggest advantage of speech based biometrics is the ability to perform authentication where a direct physical or visual contact with the subject is not feasible. In 2011, a new alternative speech production mode [10] named Whisper is used by subjects in natural conversation to protect the privacy. Due to the profound differences between whisper and neutral speech in both excitation and vocal tract function, the performance of speaker identification systems trained with neutral speech degrades significantly. In this work, a seamless neutral/whisper mismatched closed-set speaker recognition system is developed. First, performance characteristics of a neutral trained closed-set speaker ID system based on a Mel-frequency cepstral coefficient–Gaussian mixture model (MFCC-GMM) framework is considered. It is observed that for whisper speaker recognition, performance degradation is concentrated for only a subset of speakers. Next, it is shown that the performance loss for speaker identification in neutral/whisper mismatched conditions is focused on phonemes other than low-energy unvoiced consonants. Finally, a system for seamless neutral/whisper speaker identification is proposed, resulting in an absolute improvement of 8.85%–10.30% for speaker recognition; with the best closed set speaker ID performance of 88.35% obtained for a total of 961 read whisper test utterances, and 83.84% using a total of 495 spontaneous whisper test utterances. Whispered speech is a natural mode of speech production that may be employed in public situations to protect the content of speech information. When speaking on a cell phone in a public setting, a speaker may prefer to whisper when providing their credit card number, bank account number, or other personal information. Whispered speech production and acoustic characteristics: In neutral speech, voiced phonemes are produced through a periodic vibration of the vocal folds, which regulates air flow into the pharynx and oral cavities. However, for whispered speech, the shape of the pharynx is adjusted such that the vocal folds do not vibrate, resulting in a continuous air stream without periodicity. Changes in focal fold physiology due to functional voice disorders, trauma, or disease can cause changes to speech production which can often appear to take on whisper speech characteristics. 

Phoneme analysis for SPEAKER ID with WHISPERED SPEECH: The acoustic characteristics of vowels and voiced consonants differ significantly between whispered and neutral speech, while the spectral properties of unvoiced consonants are relatively similar between whispered and neutral speech. It is noted that, due to the absence of periodic excitation, the voiced consonants in whispered speech are similar to the unvoiced consonants.

Finally to conclude, whisper is an alternative speech production mode that is employed by individuals for communication in public circumstances to protect personal privacy. However, the performance of traditional MFCC-GMM speaker ID systems degrades rapidly due to the significant differences in speech production between whispered and neutral speech. The goal of this study therefore has been to develop a seamless closed-set speaker recognition system that works for both neutral/whisper mismatched and neutral/neutral matched scenarios. In 2010, the task of text-independent speaker verification [11]: given a sample from a speaker and a claimed identity that need to decide whether the claim is true or false was proposed. Prosody, the intonation, rhythm and stress patterns in speech, is not directly reflected in the spectral features and, hence, a system based on prosodic information should be highly independent from a low level spectral system. That systems based on prosodic information can lead to significant improvements when combined with a state-of-the-art low-level system. Prosodic information has been successfully used for speaker recognition for more than a decade. The best-performing prosodic system to date has been one based on features extracted over syllables obtained automatically from speech recognition output. The features are then transformed using a Fisher kernel, and speaker models are trained using support vector machines (SVMs).

PROSODIC FEATURES: Syllables are estimated automatically using the output of an automatic speech recognition (ASR) system, and more than a hundred measurements based on pitch and energy signals, along with the duration of the syllable and its constituents (onset, nucleus, and coda) are extracted over each syllable. These features called syllable-based NERF’s (non-uniform extraction region features), or SNERFs. In this work focus is on comparing modeling methods for the simpler set of features, which is called energy valley-based polynomial approximation (EV-PA). Essentially, the speech signal is segmented into regions by splitting the voiced regions wherever the energy signal reaches a local minimum. Finally this work presented study of two different modeling methods, JFA modeling of GMM means and SVM modeling of GMM weights, for a subset of simple prosodic features obtained by polynomial approximations of the pitch and energy signals over pseudo-syllables. Results indicate that, for these features, the JFA method greatly outperforms the SVM method, and that the combination of both methods leads to significant gains over JFA alone.
Our results extend the previous use of JFA for these features, including the modeling of different sequence lengths and different polynomial order approximations. It demonstrates a gain of 25\% on the JFA method after these additions. In 2010, Two important components of a speaker identification system proposed are: the feature extraction and the classification tasks\[12\]. First, features must be robust to noise and they must also be able to provide discriminating information that the classifier can use to determine the speaker's identity. Second, the classifier must take the features that have been extracted from a sentence and label them as corresponding to one of the enrolled speakers. Results using the King database shows that both fusion methods lead to enhanced performance. Speaker recognition is the concept of using a machine that is capable of identifying an individual by the spectral properties of their voice. Speaker recognition systems typically operate in two types of modes: verification and identification. Verification will validate a person's identity by comparing the captured speech to its own biometric templates that have been saved in the database, whereas the identification mode will search templates of all the users in the database for a match. This is main consideration of this work. In order for a speaker identification system to be effective, it must be robust to changes in channel distortion and the distortion may change with time. Therefore, robust features are sought for speaker identification that is not easily affected by changes in the channel. Feature Extraction: A. Frame Selection: The speech is sampled at 8 kHz and the features are processed in frame sizes of 240 samples with 160 samples overlap between frames. Each frame has a Hamming window applied. Energy thresholding is performed over all frames of a sentence to determine the relatively high energy speech frames. These high energy frames are further reduced to have at least six roots of the linear prediction polynomial (of order 12) meet the following two criteria, namely, (1) Angles between 300 Hz and 3700 Hz and (2) Magnitudes greater than or equal to 0.88. B. Mean Removed Mel Frequency Cepstrum (MRMFCC). The Mel cepstrum exploits auditory as well a decorrelation property in the cepstrum. The DCT of the output of the mel filter bank is referred to as the Mel cepstrum. The mean of all the frames, not just the high energy frames, is computed and removed from the selected high energy frames before being classified using the Vector Quantizer. This feature is referred to as the mean removed Mel frequency cepstrum (MRMFCC). C. Pole Filtered Mean Removed Cepstrum (PFMRCEP).

The PFMRCEP feature is used to remove any channel distortion that may be present in the speech signal which will corrupt the features and be detrimental to the classification performance of the identification system. Before the PFMRCEP can be computed, the linear prediction cepstrum clp (n) must be computed. Finally, the PFMRCW is a more robust feature than the MRACW as observed over all sessions of the King database. This suggests that the pole filtering method gets a better channel estimation than the mean of all the feature vectors. The PFMRCW was also a better feature than PFMRCEP when tested on sessions 6-10 of the King database indicating that PFMRCW is more robust to severe changes in the channel distortion than PFMRCEP.

II. CONCLUSION

Many research works are carrying out in the field of speaker recognition in current era. As compared with many of the different categories of biometrics the speaker recognition found quite challenging and combo of inevitable issues because of variation in tone and fluctuation of speech of human who is already authenticated. Here it is planned to survey some of the articles. While the survey, many of the issues and challenges are found in speaker recognition and simultaneously came across with various speaker recognition types and feature extraction methods.

III. FUTURE DIRECTIONS

Although many recent advances and successes in speaker recognition have been achieved, there are still many problems for which good solutions remain to be found. Most of these problems arise from variability, including speaker-generated variability and variability in channel and recording conditions. It is very important to investigate feature parameters that are stable over time, insensitive to the variation of speaking manner, including the speaking rate and level and robust against variations in voice quality due to causes such as voice disguise or colds. It is also important to develop a method to cope with the problem of distortion due to telephone sets and channels, and background and channel noises. From the human-interface point of view, it is important to consider how the users should be prompted, and how recognition errors should be handled. Studies on ways to automatically extract the speech periods of each person separately from a dialogue involving more than two people have recently appeared as an extension of speaker recognition technology.
REFERENCES


