A Review On Speaker Recognition Approaches And Challenges

Varun Sharma
M. Tech (Student)
MMU,Solan (HP)

Dr. P K Bansal
Director
MMU,Solan (HP)

Abstract

Speech recognition is the ability to identify spoken words, and speaker recognition is the ability to identify who is saying them. The usefulness of identifying a person from the characteristics of his voice is increasing with the growing importance of automatic information processing and telecommunications. In this paper we provide a brief overview of the area of speaker recognition, describing its system, various modules of feature extraction and modelling, applications, underlying techniques and some indications of performance. Following this overview we will discuss some of the strengths and weaknesses of current speaker recognition technologies and outline some potential future trends in research, development and applications. The paper concludes with discussions on future trends and research opportunities in this area.

IJERT

1. Introduction

The human ear is a marvelous organ. Beyond our unique human ability to receive and decode spoken language, the ear supplies us with the ability to perform many diverse functions. These include, for example, localization of objects, enjoyment of music, and the identification of people by their voices. Currently, along with efforts to develop computer procedures that understand spoken messages, there is also considerable interest in developing procedures that identify people from their voices[1].Being able to speak to your personal computer, and have it recognize and understand what you say, would provide a comfortable and natural form of communication. It would reduce the amount of work you have to do leaving your hands free. It would also help in some cases if the computer could tell who was speaking [2].The speech signal conveys information at different levels to the listener. At the primary level, speech conveys a message via words. But at other levels speech conveys information about the language being spoken and the emotion, gender and, generally, the identity of the speaker[3].Since the computers have been evolved we are dealing with various research activities in the area of man machine interface. The input peripherals are although very popular mediums to interact with the computer but has some limitations as keyboard requires a certain amount of skill for effective and fast usage and mouse on the other hand requires a good hand and eye coordination. The physically challenged people find computers difficult to use. Speech which is a natural and very easy way of exchanging the information if used as a medium to interact with the computer and can solve all these problems. Signal processing has made it possible for computers to follow human voice commands and understand human languages as speech can be characterized in terms of signal carrying message information.

Speech signal processing could be divided into three different tasks: Analysis, Recognition and Coding. Recognition research fields could be subdivided into three parts: Speech, Speaker and Language recognition systems. While speech recognition aims at recognizing the word spoken in speech, language recognition aims at the detection of language spoken and the goal of speaker recognition systems is to extract, characterize and recognize the information in the speech signal conveying speaker identity. In automatic speech recognition, an algorithm takes the listener’s role in deciphering speech waves into the underlying textual message. In automatic speaker recognition, an algorithm generates a hypothesis concerning the speaker’s identity or authenticity [4-7]. In this paper, we concentrate ourselves on speaker recognition systems (SRS).

Automatic Speaker Recognition is the field of digital signal processing related to the recognition of the people based on their voice. No two individuals are identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on[8].It is a popular biometric identification technique used for authenticating and monitoring humans using their speech signal. The method has several benefits (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 transducers as microphones are now ubiquitous on most portable devices (cellular phones, PDAs and laptops). The key applications that are driving this demand in speaker verification/identification technology are tele-commerce and forensics where the objective is to automatically authenticate speakers of interest using his/her conversation over a voice channel (telephone or wireless phone)[9]. This technique makes it possible to use the speaker's voice to verify their identity and control access to various services.

2. SPEAKER RECOGNITION

Speaker recognition is the process of automatically recognizing who is speaking Speaker recognition system is often classified into closed-set recognition and open-set recognition. Just as their names suggest, the closed-set refers to the cases that the unknown voice must come from a set of known speakers; and the open-set means unknown voice may come from unregistered speakers, in which case we could add 'none of the above' option to this identification system. Moreover in practice speaker recognition systems could also be divided according to the speech modalities: text-dependent recognition, text-independent recognition. For text dependent SRS, speakers are only allowed to say some specific sentences or words, which are known to the system. On the contrary, as for the text-independent SRS, they could process freely spoken speech, which is either user selected phrase or conversational speech. Compared with text-dependent SRS, text-independent SRS are more flexible, but more complicated [10]. All above speaker recognition techniques can be classified into speaker verification and identification. Speaker verification is the process of accepting or rejecting the identity claim of speaker while identification is the process of determining which registered speaker provides a given utterance [11]. Both classifications discussed as follows in detail.

2.1 Speaker Verification

Speaker verification is used to determine whether a person claims to be according to his/her voice sample. This task is also known as voice verification and speaker detection. Speaker verification is a 1:1 match where one speaker's voice is matched to one template (also called a "voice print" or "voice model") or in other sense Pattern Matching between the claimed speaker model registered in the database and the imposter model will be performed then (see figure 1). If the match is above a certain threshold, the identity claim is verified. Using a high threshold, system gets high safety and prevents impostors to be accepted, but in the mean while it also takes the risk of rejecting the genuine person, and vice versa

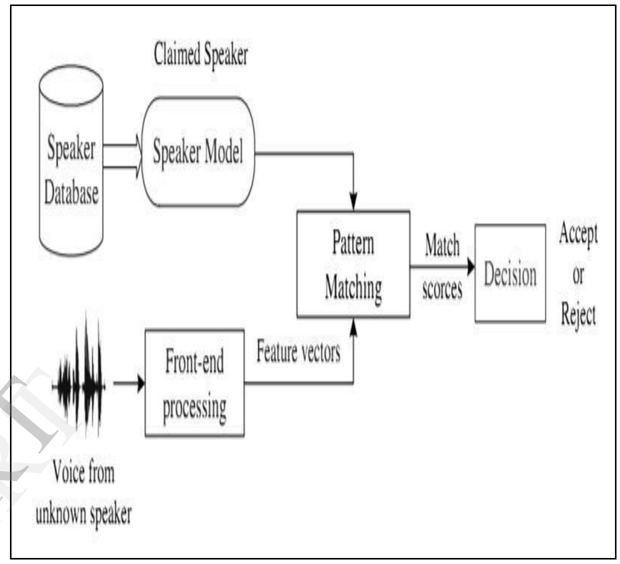


Figure 1: Speaker Verification System

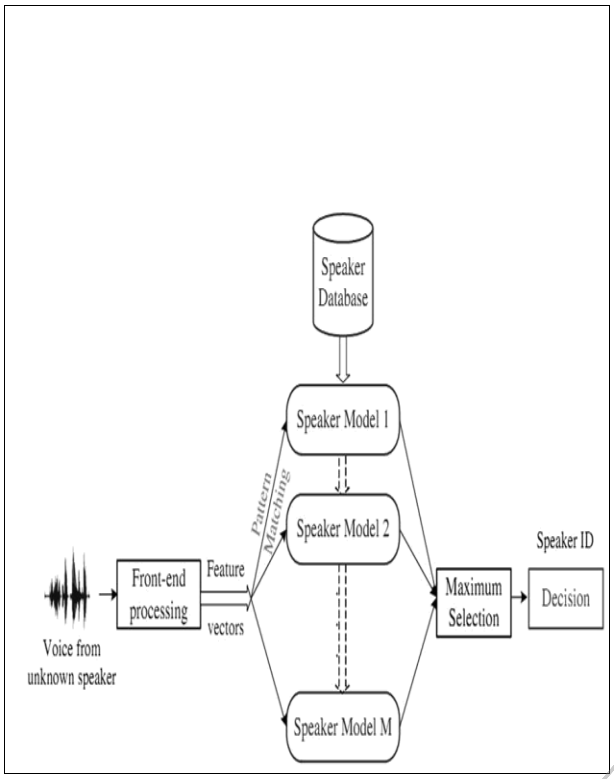
IJERT

2.2 Speaker Identification

Speaker identification is the process of finding the identity of an unknown speaker by comparing his/her voice with voices of registered speakers in the database. It is a one to- many comparisons (1: N match where the voice is compared against N templates) [12] (see Figure 2). In Speaker identification System, M speaker models are scored in parallel and the most one of the speaker's ID in the database, or will be 'none of the above' if and only if the matching score is below some threshold and it's in the case of a open most-likely one is reported, and consequently decision will be open-set Speaker identification system. The desirable features for SIS should possess the following attributes: [13]

- Easy to extract, easy to measure, occur frequently and naturally in speech
- Not be affected by speaker physical state (e.g. illness)
- Not change over time, and utterance variations (fast talking vs. Slow talking rates)
- Not be affected by ambient noise
- Not subject to mimicry

Figure 2: Speaker Identification System



The speaker ID problem may further be subdivided into closed set and open set. The closed set refers to a case where the speaker is known and belongs to a set of M speakers. In the open set case, the speaker may be out of the set and hence, a "none of the above" category is necessary. Another distinguishing aspect of speaker ID systems is that they can either be text-independent or text-dependent depending on the application. In the text-independent case, there is no restriction on the sentence or phrase to be spoken, whereas in the text-dependent case, the input sentence or phrase is fixed for each speaker. A text-dependent is commonly used in speaker verification systems in which a person's password is critical for verifying his/her identity.

IJERT

3. Speaker Recognition System

Speaker Recognition system makes it possible to use the speaker's voice to verify their identity and control access of the desired services. Speaker recognition system is having three main components: Front -end Processing or Feature Extraction, Speaker Modeling, Pattern Matching or Logical decision (see Figure 3). To get the feature vectors of Incoming voice, front end processing will be performed. Feature vectors are used to create a speaker model. The pattern matching is responsible for comparing the features to speaker models. The decision module analyzes the similarity score (statistical or deterministic) to make a decision. Using a high threshold, system gets high safety and prevents impostors to be accepted, but in the mean while it also takes the risk of rejecting the genuine person, and vice versa.

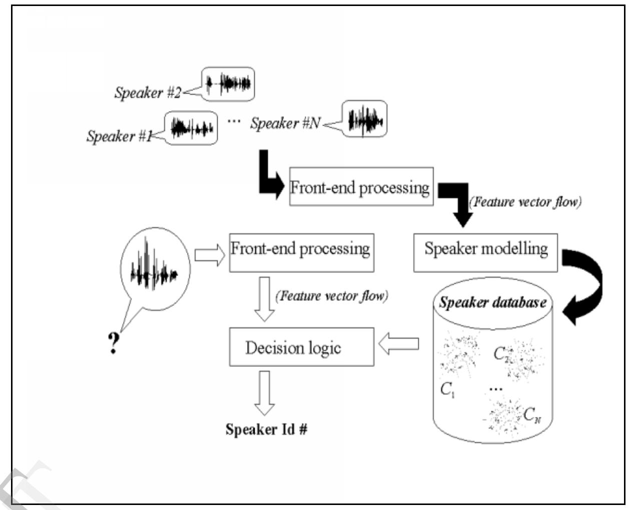


Figure 3: Speaker Recognition System

3.1 Front End Processing

It transforms the high dimensional speech signal to a relatively low dimensional features subspace while preserving the speaker discriminative information. The speaker's voice signal contains both the features of speech and the speaker personality characteristics. And also the speaker's model is not derived from the speech signal, but received by extracting features from the speech signal. In other words, the model is a speech feature model of speaker. The test tone is compared and matched to the speaker's model only after features parameter extracted, and the training speech get the model after feature extracted so feature extraction is important part in speaker recognition system. It also computes the performance of the recognition system

Feature extraction process includes a sampling, a quantization, a pre – emphasis, a windowing, and a feature extraction.

3.1.1 Pre- emphasis

Pre-emphasis are necessary because, voiced sections of speech signal naturally have a negative spectral slope (attenuation) of approximately 20db per decade due to physiological characteristics of speech production. So high frequency formants have small amplitude with respect to low frequency formants mean the energy in the signal decreases as the frequency increases. Pre-emphasis increases the energy in the parts of the signal by the amount inversely proportional to its frequency. Thus, as the frequency increases, pre-emphasis raise the energy of the speech signal by an increasing amount [14].

3.1.2 Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n)$, $0 \leq n \leq N -1$, where N is the number of samples in each frame, then the result of windowing of signal is y(n) = x(n) * w(n) where x(n) is the speech signal being processed. Windowing is followed by framing, where the speech signal is made stationary by dividing it into overlapping fixed duration segments called frames which can capture the speaker specific characteristics[15].

3.1.3 Feature Extraction

It is a process to extract speaker's personal feature traits. There are several feature extractors popular in literature such as Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel Cepstral Coefficients (MFCC), Perceptual Linear Predictive Cestrum Coefficients (PLPCC), and Real Cepstral Coefficients (RCC)

.

IJERT

Linear Predictive Coefficients (LPC): LPC is defined as a digital method for encoding an analog signal in which a particular value is predicted by a linear function of the past values of the signal. Human speech is produced in the vocal tract which can be approximated as a variable diameter tube. The linear predictive coding (LPC) model is based on a mathematical approximation of the vocal tract represented by this tube of a varying diameter. At any particular time, t, the speech sample s (t) is represented as a linear sum of the p previous samples. The most important aspect of LPC is the linear predictive filter which allows the value of the next sample to be determined by a linear combination of previous samples. Linear predictive coding may reduce bit rate significantly and at this reduced rate the speech has a distinctive synthetic sound and there is a noticeable loss of quality. However, the speech is still audible and it can still be easily understood. Since there is information loss in linear predictive coding, it is a lossy form of compression. LPC does not represent the vocal tract characteristics from the glottal dynamics and also is takes more time and computational cost to create the model of each speaker. Linear Predictive Coding is an analysis/synthesis technique to lossy speech compression that attempts to model the human production of sound instead of transmitting an estimate of the sound wave.

Linear Predictive Cepstral Coefficients (LPCC): LPCC reveals the differences of the biological structure of human vocal track and is computed through recursion from the LPC Parameters to the LPC cepstrum according to an all pole model At the core of the LPCC feature extraction algorithm is the Linear Prediction Coding (LPC) technique [16, 17] which assumes that any speech signal can be modeled by a linear source-filter model. This model assumes two sources of human vocal sounds: the glottal pulse generator and the random noise. The glottal pulse generator creates voiced sounds. This source generates one of the measurable attributes used in voice analysis: the pitch period. The random noise generator produces the unvoiced sounds and the vocal tract serves as the filter of the model that produces intensification at specific formants. In LPCC feature extraction, the filter is typically chosen to be an all-pole filter. The parameters of the all-pole filter are estimated using an auto-regressive procedure where the signal at each time instant can be determined using a certain number of preceding samples.

Mathematically this can be expressed as

$$S\ (t) = \sum_{i=1}^{p} ais(t-i) + e(t)$$

Where s(t) is the speech signal at time instant t is determined by p past samples s(t-i) where i represents the discrete time delay. e (t) is known as the excitation term (random noise or glottal pulse generator) which also signifies the estimation error for the linear prediction process and *ai* denotes the LPC coefficients.

Mel Cepstral Coefficients (MFCC): MFCC's are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. It is a technique based on hearing behavior that cannot recognize frequencies over 1KHz. The signal is expresses in the MEL scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz [18, 19]. Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. MFCC features are based on the known variation of the human ear's critical bandwidths with frequency.

Perceptual Linear Predictive Cestrum Coefficients (PLPCC) : Perceptual Linear Predictive Cepstral Coefficients (PLPCC) are based on the magnitude spectrum of the speech analysis window. Unlike MFCC and LPC which are cepstral methods, the PLPCC is a temporal method and models the speech auditory spectrum through a low order all pole model. Revathi [20] details the steps followed to calculate the coefficients of the PLPCC. First, compute the power spectrum of a windowed speech. Second, group the results to 23 critical bands using bark scaling for sampling frequency of 8 kHz. Third, perform loudness equalization and cube root compression to simulate the power law of hearing. Fourth, perform inverse Fast Fourier Transform (IFFT). Fifth, perform LP analysis by Levinson- Durbin algorithm [21]. Lastly, convert LP coefficients into cepstral coefficients. The relationship between frequency in Bark and frequency in Hz is specified as in f(bark) = 6*arcsin h(f(Hz)/600)

Real Cepstral Coefficients (RCC): Rasta filtering takes advantage of the fact that the rate of change of nonlinguistic components in speech often lies outside the typical rate of change of the vocal-tract shape. Therefore, it suppresses the spectral components that change either more slowly or more quickly than the typical rate of change of speech. The RASTA approach can be combined with the PLPCC method to get the low-pass transfer function H(z) [22]. RASTA filtering reduces the accuracy of the system in the absence of noise. However it increases its accuracy significantly in the presence of severe noise .

IJERT

3.2 Speaker Modeling

During enrollment, speech from a speaker is passed through the feature extraction module and the feature vectors are used to create a speaker model. Desirable attributes of a speaker model are: (1) a theoretical underpinning so one can understand model behavior and mathematically approach extensions and improvements; (2) generalizable to new data so that the model does not over fit the enrollment data and can match new data; (3) parsimonious representation in both size and computation. There are many modeling techniques that have some or all of these attributes and have been used in speaker verification systems. The selection of modeling is largely dependent on the type of speech to be used, the expected performance, the ease of training and updating, and storage and computation considerations. A brief description of some of the more prevalent modeling techniques is given next.

3.2.1 Template Matching

In this technique, the model consists of a template that is a sequence of feature vectors from a fixed phrase. During verification a match score is produced by using dynamic time warping (DTW) to align and measure the similarity between the test phrase and the speaker template. This approach is used almost exclusively for text-dependent applications.

3.2.2 Nearest Neighbor

In this technique, no explicit model is used; instead all features vectors from the enrollment speech are retained to represent the speaker. During verification, the match score is computed as the cumulated distance of each test feature vector to its k nearest neighbors in the speaker's training vectors. To limit storage and computation, feature vector pruning techniques are usually applied.

3.2.3 Neural Networks

The particular model used in this technique can have many forms, such as multi-layer perceptions or radial basis functions. The main difference with the other approaches described is that these models are explicitly trained to discriminate between the speaker being modeled and some alternative speakers. Training can be computationally expensive and models are sometimes not generalizable.

3.2.4 Hidden Markov Models

This technique uses HMMs, which encode the temporal evolution of the features and efficiently model statistical variation of the features, to provide a statistical representation of how a speaker produces sounds. During enrollment HMM parameters are estimated from the speech using established automatic algorithms. During verification, the likelihood of the test feature sequence is computed against the speaker's HMMs. For text-dependent applications, whole phrases or phonemes may be modeled using multi-state left-to right HMMs. For text-independent applications, single state HMMs, also known as Gaussian Mixture Models (GMMs), are used.

3.3 Logical Decision

It is the process of pattern matching which is the actual comparison of the extracted features with known speaker models (or templates), this result in a matching score which quantifies the similarity in between the voice recording and a known speaker model. Pattern matching is often based on Hidden Markov Model (HMMs), a statistical model which takes into account the underlying variations and temporal changes of the acoustic pattern. Alternatively Dynamic Time

Warping (DTW) is used; this algorithm measures the similarity in between two sequences that vary in speed or time, even if this variation is non-linear such as when the speaking speed changes during the sequence [23].

4. APPLICATIONS

Speaker recognition has its vast applications in various fields. Following is the list of few technologies where speaker recognition has been or is currently used.

Access Control: Here speaker recognition can be used as a very effective biometric authentication. This includes controlling the access for computer network or websites required some subscription. This May be used for automated password reset services.

Forensic Science: Includes legal use of acoustic samples. Here by the use of forensic speaker comparison, one may conclude whether the two speech samples are from the same speaker or not. This introduces the concept of ear witness in many of the legal cases.

Transaction Authentication: For telephone banking, in addition to password account access control, higher levels of verification can be used for more sensitive transactions. More recent applications are in user verification for remote electronic and mobile purchases (e-and m-commerce).

Law Enforcement: Some applications are home-parole monitoring (call parolees at random times to verify they are at home) and prison call monitoring (validate inmate prior to outbound call).This Includes surveillance and speaker tracking.

Speech Data Management: In voice mail browsing or intelligent answering machines, use speaker recognition to label incoming voice mail with speaker name for browsing and/or action (personal reply).

Personalization: In voice-web or device customization, store and retrieve personal setting based on user verification for multi-user site or device. There is also interest in using recognition techniques for directed advertisement or services, where, for example, repeat users could be recognized or advertisements focused based on recognition of broad speaker characteristics (e.g. gender or age).

5. FACTOR AFFECTING SPEAKER RECOGNITION

The performance of the speaker recognitions are affected by many factors simultaneously. The interpretation is in details that were driving the speaker recognition system and affect its performance described as following [24]:

Noise- The background noise is the most significant factor for the speaker recognition accuracy, which is high for the clean samples but deteriorates quickly for noisy samples. Only babble noise has no significant influence.

Microphones- Results were best without mismatch the microphone quality itself is insignificant and it has not much more impact on the efficiency of speaker recognition systems

Disguise- Deliberate cheating is possible, the recognition fails in most of the cases but it has no much more impact on speaker recognition systems.

Quality of the Voice Sample- For clean voice samples, the higher quality of microphones leads to the better results. However, ordinary quality of microphone gives almost the same results as good quality microphones.

Voice Sample Length- In general it was assumed that longer samples improve the speaker recognition efficiency but could not be verified. There is no significant difference to clean samples. In background noise the short samples provide better results but the difference is within the confidence level.

Language used in training and testing- There is no much more impact on the efficiency of speaker recognition in native language speech. For the noisy samples, the English language samples give better results.

Speech modality: Whether the system is text-dependent or text–independent. The role of text is insignificant however text independent systems are more complex to design.

Speaker population: This has a very significant impact on the efficiency of the system. As the number of speaker increases in our database, we have to compromise with the correct recognition of the speaker.

6. STRENGTHS AND WEAKNESSES

Speaker Recognition has its various strengths and weaknesses and the following criteria are used to evaluate the suitability of speaker recognition as biometrics.

IJERT

Collectability: Voice recordings are easy to obtain and do not require expensive hardware. The real advantage of voice recognition is that it can be done over telephone lines or using computer microphones, with variable recording and transmission quality. Pattern matching algorithms must be able to handle ambient noise and differing quality of the recordings.

Portability: Speaker verification is easy to use, has low computation requirements (can be ported to cards and handhelds) and, given appropriate constraints, has high accuracy.

Acceptability: Speaker recognition is unobtrusive; speaking is a natural process so no unusual actions are required. When speaker recognition is used for surveillance applications or in general when the subject is not aware of it then the common privacy concerns of identifying unaware subjects apply. Moreover speaker information can be obtained easily from almost anywhere using the familiar telephone network (or internet) with no special user equipment or training.

Circumvention: A major issue with speaker recognition is spoofing using voice recordings. The risk of spoofing with voice recordings can be mitigated if the system requests a random generated phrase to be repeated, an impostor cannot anticipate the random phrase that will be required and therefore cannot attempt a playback spoofing attack.

IJERT

Performance: Robustness is very dependent on the setup, when telephone lines or computer microphones are used the algorithms will have to compensate for noise and issues with room acoustics. Furthermore speaker recognition is, because the voice is a behavioral biometric, impacted by errors of the individual such as misreading and mispronunciations.

Mobility: Mobility of system means that people are using verification systems from more uncontrolled and harsh acoustic environments (cars, crowded airports), which can stress accuracy.

Variability: The varied microphones and channels that people use can cause difficulties since most speaker verification systems rely on low-level spectrum features susceptible to transducer/channel effects.

Universality: Obviously for people who are mute or having problems with their voice due to severe illness this biometric solution is not useable.

Permanence: Speech signal used for speaker recognition is a behavioral signal that may not be consistently reproduced by a speaker so an issue with speaker recognition is that the voice changes with ageing, and is also influenced by factors such as sickness, tiredness, stress, etc.

7. CONCLUSION AND FUTURE TRENDS

In this paper, we have presented an overview of speaker recognition system which includes various methods of feature extraction and feature modeling. Focus is also flashed on the various applications and factors affecting the system. The speaker recognition system still has various drawbacks which can be further reduced by carrying out research in sub-domains and merging of other biometrics system with speaker recognition. The main application for the technology is in the area of access control, where the speakers are required to be authenticated before they can be allowed access to certain facilities or some other restricted services in various domains which is having some secured information. The future trend in access control is to integrate speaker verification technology into a multi-level and a hybrid authentication approach, where results from different biometric technology like finger print, face, iris and speaker recognition could be fused together to achieve better reliability in authentication. However, the biggest advantage of speech based biometrics is the ability perform authentication where a direct physical or visual contact with the subject is not feasible. Thus the technology has a clear advantage for authenticating transactions that occur over the voice channel like telebanking. A more controversial application of speaker verification technology is in the area of forensics where the results of the technique could be offered as evidence in judicial trials. Compared to finger-printing and DNA based authentication technology, the existing speaker verification techniques have their drawbacks and limitations due their sensitivity to corruption by noise and the ability to masquerade the signal using voice recording devices. We believe that there is an enormous potential for speaker verification and recognition technology in multimedia and biometric applications. However, key challenges still remain to be solved and are currently limiting the wide-scale deployment of the technology. These challenges motivate further research and investment in some of the following important directions:

Exploitation of higher-levels of information: In addition to the low-level spectrum features used by current systems, there are many other sources of speaker information in the speech signal that can be used. These include idiolect (word usage), prosodic measures and other long-term signal measures. This work will be aided by the increasing use of reliable speech recognition systems for speaker recognition R&D. High-level features not only offer the potential to improve accuracy, they may also help improve robustness since they should be less susceptible to channel effects.

Focus on real world robustness: Speaker recognition continues to be data-driven field, setting the lead among other biometrics in conducting benchmark evaluations and research on realistic data. The continued ease of collecting and making available speech from real applications means that researchers can focus on more real-world robustness issues that appear. Obtaining speech from a wide variety of handsets, channels and acoustic environments will allow examination of problem cases and development and application of new or improved compensation techniques.

Emphasis on unconstrained tasks: With text-dependent systems making commercial headway, R&D effort will shift to the more difficult issues in unconstrained situations. This includes variable channels and noise conditions, text-independent speech and the tasks of speaker segmentation and indexing of multi-speaker speech.

8. ACKNOWLEDGMENTS

We are thankful to all researchers and authors of various manuscripts from where we got valuable information for completing the work.

9. REFERENCES

[1] George R Doddington, Member, IEEE, “Speaker Recognition- Identifying People by their Voices”, proceedings of the IEEE, Vol. 73, no. 11, pp. 1651-1664, November 1985

[2] Richard D. Peacocke, and Daryl H. Graf, “An Introduction to Speech and Speaker Recognition “, IEEE, pp. 26-33, August 1990.

[3] Douglas A. Reynolds, “An Overview of Automatic Speaker Recognition Technology”, IEEE, pp. IV – 4072 - IV – 4075, 2002.

[4] Atal, B. S., “Automatic recognition of speaker from their voices”, Proc IEEE 64, pp. 460-475, 1976.

[5] A.E. Rosenberg, “Automatic Speaker verification: A review”, Proc IEEE 64, pp. 475-487, 1976.

[6] A.E. Rosenberg, F.K. Soong, “Recent research in automatic speaker recognition”, Advances in Speech Signal Processing, Marcel Dekker, New York, pp. 701–738, 1991.

[7] J.P. Campbell, “Speaker recognition: A tutorial”, Proc. IEEE 85, pp. 1437–1462, 1997

[8] Tomi kinnunen, Haizhou Li, “An overview of text-independent speaker recognition: From features to supervectors”, Speech Communication 52, pp. 12-40, 2010.

[9] Amin Fazel, Shantanu Chakrabartty, “An Overview of Statistical Pattern Recognition Techniques for Speakeer verification”, IEEE Circuit and System Magzine, pp. 62-81, 2011.

[10] Ling Feng , Kgs. Lyngby “Speaker Recognition”, Thesis, Technical University of Denmark Informatics and Mathematical Modeling, Denmark, 2004

[11] Minh, Do No, “An Automatic Speaker Recognition System”, White paper, Digital Signal Processing Mini-Project, Audio Visual Communications Laboratory, Swiss Federal Institute of Technology, Lausanne, , Switzerland, 1996 pp.1-14.

[12] J. A. Markowitz and colleagues, “J. Markowitz, Consultants”, http://www.jmarkowitz.com/glossary.html.

[13] Ing. Milan Sigmund, CSc. “Speaker Recognition, Identifying People by their Voices”, Brno University of Technology, Czech Republic, Habilitation Thesis, 2000

[14] HAI-YAN YANG,XIN-XING JING, “Performance test of parameters for speaker recognition based on SVM-VQ”, proc of International conference on Machine Learning and Cybernetics, july, 2012

[15] J. Wu and J. Yu, "An improved arithmetic of MFCC in speech recognition system," in Electronics, Communications and Control (ICECC), 2011 International Conference on, 2011, pp. 719-722.

[16] J. Makhoul, “Linear prediction: A tutorial review,” *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[17] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[18] José Ramón Calvo de Lara, “A Method of Automatic Speaker Recognition Using Cepstral Features and Vectorial Quantization”, M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 146 – 153, 2005. Springer- Verlag Berlin Heidelberg 2005.

[19] P. Chakraborty1, F. Ahmed , Md. Monirul Kabir , Md. Shahjahan1, and Kazuyuki Murase “An Automatic Speaker Recognition System”, M. Ishikawa et al. (Eds.): ICONIP 2007, Part I, LNCS 4984, pp. 517–526, 2008. Springer-Verlag Berlin Heidelberg 2008.

[20] A. Revathi and Y. Venkataramani, "Text Independent Composite Speaker Identification/Verification Using Multiple Features," in Computer Science and Information Engineering, 2009 WRI World Congress on, 2009, pp. 257-261.

[21] P. Delsarte and Y. Genin, "The split Levinson algorithm," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 34, pp. 470- 478, 1986.

[22] R. J. Mammone, Z. Xiaoyu, and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," Signal Processing Magazine, IEEE, vol. 13, p. 58, 1996.

[23] “Speaker Recognition”, http://www.biometric-solutions.com/solutions/index.php?story=speaker_recognition

[24] Satyanand Singh, Dr. E. G. Ranjan, “MFCC VQ based Speaker Recognition and its Accuracy Affecting Factors”, International Journal of Computer Applications, Volume 21, no. 6, May 2011

IJERT