A Reliable Speaker Verification System Based on LPCC and DTW

Rekha Nair and Nirmala Salam
Biometrics Division, CDAC, Mumbai, India

***Abstract*— Human voice can serve as a password/key for access to various services. This voice is used for verifying speaker in speaker verification system based on the features extracted from the voice signal. In automated speaker verification the speaker's voice signal is processed to extract speaker-specific information which is used to generate voiceprint also known as a template that cannot be replicated by any source except the original speaker. An automatic speaker verification system based on LPCC (Linear Predictive Cepstral Coefficient) and DTW (Dynamic Time Warping) has been proposed in this paper which is based on the assumption that the shape of the vocal tract governs the nature of the sound being produced.**

***Keywords*— Speaker Verification, LPC, LPCC, DTW**

I. INTRODUCTION

For a human, speaking is naturalness, so embedding speaker verification technology into applications is non intrusive from the user's viewpoint, which makes speaker verification the most obvious application of any biometric authentication or verification technique [1]. The main advantage of using voice as a person authentication technology is that it doesn't require any special or costly equipment. Only a microphone or any standard telephone, mobile phone can suffice the requirement of voice capturing device for speaker verification which is contrast to other authentication technique which are image based and uses more sophisticated and costly hardware such as iris scanner and fingerprint scanner. Reynolds et al. [2] compared the speaker recognition with various methods of biometric authentication and also compared its strength and weakness. Some of the real world applications where speaker verification can be useful are secure identity management system for e-commerce applications based on voice verification, Attendance system, Mobile banking with natural language voice interface, Forensics etc**.**

Each speaker verification system has two phases: Enrolment and verification. During enrolment, the speaker's voice is recorded and a number of features are extracted to form a voice print, template, or model. In the verification phase, a speech sample or "utterance" is compared against a previously created voice print.

Speaker recognition systems fall into two categories: text-dependent and text-independent. If the text must be the same for enrolment and verification this is called text-dependent recognition. In a text-dependent system, prompts can either be common across all speakers (e.g.: a common pass phrase) or unique. In addition, the use of shared-secrets (e.g.: passwords and PINs) or knowledge-based information can be employed in order to create a multi-factor authentication scenario.

Text-independent systems are most often used for speaker identification as they require very little if any cooperation by the speaker. In this case the text during enrolment and test is different. In fact, the enrolment may happen without the user's knowledge, as in the case for many forensic applications. As text-independent technologies do not compare what was said at enrolment and verification, verification applications tend to also employ speech recognition to determine what the user is saying at the point of authentication [3].

Voice is a signal having a lot of information about the speaker. Extraction of different types of information from the voice signal is the most critical part of a speaker verification system. Before Feature extraction the voice signal is pre-processed to remove unwanted noise. Many features and extraction techniques have been proposed in literature which consists of low level features and high level features. Low-level features are generally related to physical traits of a speaker's vocal apparatus. Low level features are less error prone and provide much accurate information of the speaker. Example of low level features are pitch, volume etc. High-level features are generally related to a speaker's learned habits and style.

Higher-level features can provide useful information about a speaker, such as topic being discussed, whether the speaker is emotional, how a speaker is interacting with another speaker and so on. Higher-level information can significantly improve performance when combined with lower-level cepstral information. Examples of higher-level features include phonetic, prosodic, and lexical observations [3].

A voice signal can be fragmented in two parts: the source part and the system part. The system part consists of the smooth envelope of the power spectrum and is represented in the form of cepstrum coefficients, which can be computed by using either the linear prediction analysis (LPC) or the Mel filter-bank analysis. Most of the automatic speaker recognition systems reported in the literature utilize the system information in the form of cepstral coefficients [2] [4] [5]. These systems perform reasonably well. The LPC and LPC-Derived features have also been successfully used in the extraction of features for long. Benesty et al. [6] justify their use for the extraction of features for speech analysis involved in identification and verification purposes. Maruti Limkar et al have proposed a speaker verification system using MFCC and LPCC as cepstral feature and VQ and DTW as matching

978-1-4799-3975-6/14/$31.00 ©2014 IEEE

technique [7]. Many algorithms are available for speech recognition using LPCC but very few for speaker recognition. Wan-Chen et al have proposed speaker identification system using LPCC as a cepstral coefficient [8]. Mustafa Dhiaa Al-Hassani et al have used LPC derived features for designing text prompt speaker recognition system [9]. R.Kumar et al have proposed a multilingual speaker verification system based on LPC and neural network [10]. Sergey Novoselov et al proposed new State-GMM-supervector extractor for solving the problem of text-dependent speaker recognition and have achieved 44% reduction in EER [11]. Xiaojia Zhao et al have proposed a robust Speaker Identification system with speaker models trained in selected reverberant conditions, on the basis of bounded marginalization and direct masking [12].

Our proposed speaker verification system is based on LPCC features and DTW for pattern matching. We have carried out experiments using first LPC features and then LPCC features derived from LPC. The results of both the experiments along with the comparison are presented in the experimental results section.

II. LPC AND LPCC

Linear prediction is a signal processing technique that is used extensively in the analysis of speech signals. Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital speech signal in compressed form.

The fundamental idea is that a speech sample can be approximated as a linear combination of past samples. The importance of this method lies in its ability to provide extremely accurate estimates of the speech parameters, and in its relative speed of computation [9].

The LPC model, assumes that each samples (n) at time n, can be approximated by a linear sum of the p previous samples as in (1)

$$s[n] = \sum_{k=1}^{p} a[k]s[n-k] \quad (1)$$

where s[n] is an approximation of the present output, s[n−k] are past outputs, p is the prediction order; and {a[k]} are called the predictor coefficients.

The error between the actual sample and the predicted one can be expressed as (2)

$$e[n] = s[n] - \sum_{k=1}^{p} a[k]\, s[n-k] \quad (2)$$

where p represents the prediction order and fs the signal's sampling frequency, the formula (3) is used as a general rule of thumb:

$$p = \frac{fs}{1000} + \gamma \quad (3)$$

The value of γ described in the literature as a "fudge factor", is normally given as 2 or 3.

Linear Predictive Cepstral Coefficients (LPCC)

An important fact is that cepstrum can also be derived directly from the LPC parameter set. The relationship between cepstrum coefficients c_n and prediction coefficients a_k is represented in the following equations (4) and (5):

$$c_1 = a_1 \quad (4)$$

$$c_n = \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right). a_k . c_{n-k} + a_n, 1 < n \leq p \quad (5)$$

where p is a prediction order. It is usually said that the cepstrum, derived in such a way represents the "smoothed" version of the spectrum. Similar to LPC analysis, increasing the number of coefficients results in more details [9].

III. DYNAMIC TIME WARPING (DTW)

Speech is a time-dependent process. Hence the utterances of the same word will have different durations, and utterances of the same word with the same duration will differ in the middle, due to different parts of the words being spoken at different rates. As a result, efforts to recognize words by matching them to templates will give inaccurate results if there is no temporal alignment. To obtain a global distance between two speech patterns (represented as a sequence of vectors) a time alignment must be performed [13].

Dynamic Time Warping is a method which has been used in speaker and speech recognition since long time. DTW calculates an optimal match between two given sequences with certain restrictions. The similarity found by this algorithm gives a good indication of how well the sample and template match. DTW is guaranteed to find the lowest distance path through the matrix, while minimizing the amount of computation.

If D(i,j) is the global distance up to (i,j) and the local distance at (i,j) is given by d(i,j), then as per (6)

$$D(i,j) = \min [D(i-1,j-1), D(i-1,j), D(i,j-1] + d(i,j) \quad (6)$$

Given that D (1,1) = d(1,1) (this is the initial condition), we have the basis for an efficient recursive algorithm for computing D (i,j). The final global distance D (n,N) gives us the overall matching score of the template with the input. The input word is then recognized as the word corresponding to the template with the lowest matching score.

The paper is organized as follows. The next section elaborates the proposed approach of our system along with its components and working. The experimental results are discussed in Section 5. Section 6 discusses the future enhancements possible in the system with the concluding remarks.

IV. OUR APPROACH

Our proposed speaker verification system is based on LPCC features and DTW. Following are the different modules of the system:

A. Input

The algorithm begins by taking the input voice signal (wave file) and reading it. The input consists of two utterances by the same speaker. We have used real speech signals from TIMIT database. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech waveform file for each utterance.

B. Pre-processing

The basic idea behind speech preprocessing is to generate a signal with a fine structure as close as possible to that of the original speech signal as shown in Fig.1. In this step the raw voice signals are pre-processed by removing the silence (unvoiced) part. Apply Noise removal methods (Cepstral Mean Subtraction and Spectral Subtraction) [14] on the signal (voiced) part.

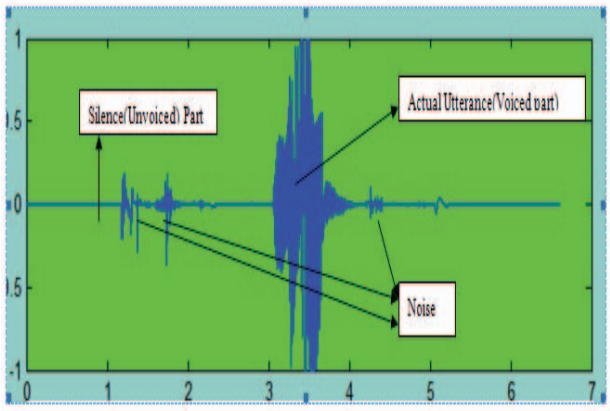

Fig 1. Voiced, Unvoiced and Silence part of a signal

C. Feature Extraction

During feature extraction, large amount of speech data is reduced into much smaller amounts which represent the important characteristics of the speech. First we extract LPC features of both the utterances. Then LPCC features are derived from these LPC features to create feature vectors for each speaker. In our system, we have carried out experiments on both LPC and LPCC features.

D. Pattern Matching

The speaker verification system works by taking the input from the speaker and comparing it with the stored template of that speaker. First, feature extraction is applied to the speaker's input sample. Then, the input feature vector is compared with the stored feature vector. The two training utterances are compared using the DTW algorithm. The distance value returned is then used to determine the threshold value to be used during verification.

During training it was noticed that distance between two utterances of the same phrase by the same speaker varied from user to user. So, it was difficult to determine a threshold value that would work for all speakers in the system. Hence, it was decided that two utterances for training should be used. Using the distance between two utterances to determine threshold values for each speaker allows the system to dynamically compute threshold value for each speaker rather than setting it statically for all speakers.

During verification, we compare the feature vector of the test signal with both the utterances of the input signal using DTW algorithm and take the average of the distances returned. If this value is less than the threshold value computed then verification is accepted, or else, verification is rejected. Block diagram of our approach is shown in Fig.2 below.

V. EXPERIMENTAL RESULTS

A. Database

Real speech signals from TIMIT database were used for experiments. TIMIT is a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects. Each transcribed element has been delineated in time. The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech waveform file for each utterance [15].

B. LPC and DTW

In the first experiment, we developed a system based on LPC features and DTW for pattern matching. During testing, it was found that the system provides a verification accuracy of approximately 92%.

C. LPCC and DTW

In the second experiment, LPCC features were derived from LPC and pattern matching was carried out using DTW. A verification accuracy of 97% was achieved during the testing phase.

D. Performance Results

Table I shows the verification accuracy results of the system. We can see that using LPCC features gives us better results than just using LPC.

Table II shows the verification accuracy of LPC and LPCC for different values of P (Prediction order).

TABLE I. VERIFICATION ACCURACY

Feature	Acceptance Accuracy	Rejection Accuracy	Overall Accuracy
LPC	91.6	91.5	92.2
LPCC	95.2	96.3	97.3

TABLE II. VERIFICATION ACCURACY FOR DIFFERENT PREDICTION ORDERS

Feature	P = 20	P = 30	P = 40
LPC	90.4	91.8	92.3
LPCC	94.8	96.3	97.1

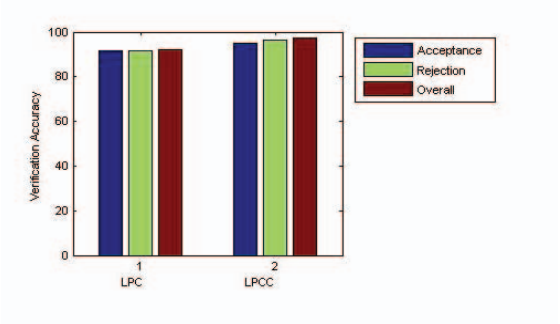

Fig. 3. Verification Accuracy for LPC and LPCC

VI. CONCLUSION AND FUTURE WORK

In this paper we have proposed a Speaker Verification system based on LPCC and DTW. The results suggest that a high recognition rate of 97% is achieved which is comparatively much better than just using LPC features. Future work will include improving channel robustness and effects of speaker ageing on the system and implementing the current system for remote speaker authentication.

REFERENCES

[1] Tomi Kinnine. Spectral features for automatic text-independent speaker recognition. Licentiates Thesis. 2003.

[2] Reynolds, L.P. Heck, "Automatic speaker recognition: current approaches and future trends," based in part on the tutorial Speaker Verification From Research to Reality, ICASSP 2001

[3] Z Saquib, N Salam, R Nair, N Pandey and A Joshi, "A survey on automatic speaker recognition systems," Signal Processing and Multimedia, vol. 123, Springer Berlin Heidelberg, pp. 134-145, 2010

[4] Jin, Q. Robust Speaker Recognition. PhD Thesis. Carnegie Mellon University 2007.

[5] S. Furui, "50 years of progress in speech and speaker recognition," Proc. SPECOM 2005, Patras, Greece, pp.1-9 (2005-10).

[6] J.Benesty, M.Sondhi and Y.Huang. Springer Handbook of Speech Processing. 2008

[7] Maruti Limkar, B.Rama Rao, Vidya Sagvekar "Speaker recognition using VQ and DTW," ICACCT 2012

[8] Wan-Chen Chen, Ching-Tang Hsieh, "Robust speaker identification system based on two-stage vector quantization," Tamkang Journal of Science and Engineering, Vol. 11, No. 4, pp. 357 366, 2008

[9] Dr. Mustafa Dhiaa Al-Hassani, Dr. Abdulkareem A. Kadhim, "Design a text-prompt speaker recognition system using LPC-derived features," ACIT2012

[10] R. Kumar, R. Ranjan, S. K.Singh, R.Kala, A. Shukla, and R. Tiwari, "Multilingual speaker recognition using neural network," Proceedings of the Frontiers of Research on Speech and Music 2009.

[11] Sergey Novoselov, Timur Pekhovsky, Andrey Shulipa1,Alexey Sholokhov, "Text-dependent GMM-JFA system for password based speaker verification," IEEE International Conference on Acoustic, Speech and Signal Processing 2014

[12] Xiaojia Zhao, Yuxuan Wang, and DeLiang Wang, "Robust speaker identification in noisy and reverberant conditions," IEEE/ACM Transactions on Audio, Speech and Language Processing, VOL. 22, NO. 4, APRIL 2014

[13] Ben Gold, Nelson Morgan. Speech and Audio Signal Processing:Processing and Perception of Speech and Music. John Wiley & Sons, Inc. 2000.

[14] Rekha Nair, Nirmala Salam, Ashutosh Singh, Ganesh Joshi, "An efficient method for additive and convolutive noise reduction," I JECSE, ISSN 2277-1956/V1N4-2078-2083, 2012

[15] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J.G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM,"*NIST*, 1993.

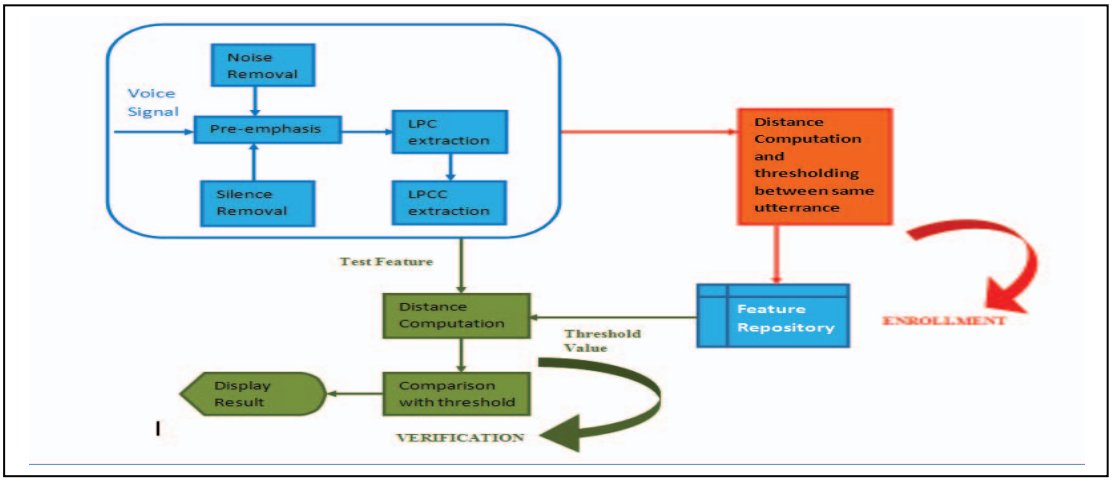

Fig.2 System Architecture of the proposed Speaker Verification System