Text-Dependent Speaker Recognition System for Telugu

Surabhi Sreekanth, Kavi Narayana Murthy
Dept. of Computer and Information Sciences
University of Hyderabad
email: knmuh@yahoo.com, surabhi_sreekanth@yahoo.co.in

Abstract

This paper is about speaker recognition. Speaker Recognition is a process of automatically recognizing who is speaking on the basis of speaker dependent features of the speech signal. Although speaker recognition is currently not as robust as other biometrics such as finger prints and retinal scans, speech has many inherent advantages and it holds promise.

In this paper we describe a system for speaker recognition designed with low security access control systems in mind. An isolated word speech recognition system is used to recognize the spoken password and then a speaker identification system is used to further confirm the identity of the user amongst a known set of users. The HTK tool kit has been used to build these systems. Hidden Markov Models based on Mel Frequency Cepstral Coefficients have been used to build models. Mahalanobis distance measure is employed.

There is little work done in speaker recognition in any of the Indian languages. In this paper, we describe a text-dependent speaker recognition system for Telugu. Experiments conducted with Telugu data have been described. Good performance has been achieved.

1 Introduction

Communication through speech is simple, natural and efficient. Input through speech does not require hand-eye coordination and is far simpler than using the keyboard, mouse and the computer screen. Speech output eliminates the need to read texts and can overcome the limitations of literacy. Speech based recognition also permits remote access.

Speaker Recognition is a process of automatically recognizing who is speaking on the basis of speaker dependent features of the speech signal without necessarily recognizing or understanding what is being spoken. A speaker recognition system must be robust against mimicking. As of today, finger prints and retinal scans are more reliable means for identification of individuals. Nevertheless, during the years ahead, it is believed that speaker recognition technology will make it possible to reliably verify the person's identity. Speech can be used for remote accessing, as through a telephone, where other methods such as finger prints and retinal scans are impractical.

Speaker Recognition can be classified into *Speaker Identification* and *Speaker Verification.* A speaker identification system finds the person who spoke the given utterance, from amongst a given set of speakers. A speaker verification system accepts or rejects the personal identity claim of a speaker. These systems can be further categorized as *text-independent* and *text-dependent.* In a text-independent system, the speaker can speak any arbitrary utterance in a particular language while in the case of text-dependent systems the speaker is required to speak known piece of text. Either all speakers may be asked to speak out the same piece of text, or, as in the case of a password based access control, users speak out their own individual passwords.

If we represent the probability of a given utterance x arising from the i^{th} speaker by $p_i(x)$, then the task of speaker verification is,

$$\begin{array}{ll} \text{if } \ p_i(x) > \mathit{Threshold} & \text{then ACCEPT} \\ & \text{else REJECT} \end{array}$$

The value of this *Threshold* can be experimentally determined.

On the other hand, a speaker identification system must decide among the N speakers in the system. The task of the speaker identification system is to find

$$\arg\max_i \{ \ \ p_i(x) \ \ \}$$

Since the system is required to make N tests and decisions, the performance of speaker identification system generally decreases as the value of N increases. While the performance of the speaker verification system is inherently independent of the number of speakers in the user database, it is still difficult to achieve high accuracy and robustness when the number of speakers is large.

There are two types of errors a speaker recognition system can make - *false acceptance* and *false rejections.* If a wrong person is accepted, then this is a false acceptance. If a right person is rejected, then it is a false rejection. In the case of speaker verification, the decision of acceptance and rejection is made based on a threshold. So, the system can be designed to reduce a particular type of error by varying threshold. For example, if the system is used for accessing sensitive information, then the value of the threshold can be changed so as to reduce the false acceptance errors. Thus, even if the false rejection errors may increase, unauthorized access can be reduced with some inconvenience to true users.

The threshold can be adjusted in such a way that the two kinds of errors occur with equal probability. This is termed as *Equal Error Rate.* This equal error rate is significantly affected by the samples in the training and test data.

Speaker recognition systems may have to work across languages. However, if it is known that the users of a system belong to particular language, then it is possible to build a system for that specific language. It has been shown that the performance of the speaker verification systems built with samples of particular language degrade when the target speakers are from a different language. This performance degradation can be restricted if the speaker models are created with a pool of training data covering many languages [2].

Today, there are around one billion people in India, speaking nearly 200 different languages. Of these 22 languages are given constitutional recognition and are considered to be major languages. A large portion of the Indian population is either illiterate or not so comfortable with using keyboard-mouse based interfaces, especially if the interactions are in English. Thus speech based interfaces have a very important role in the Indian context. Speech can empower people by helping them to overcome the barriers of language - people can simply speak the language they already know. There is no need to learn any new technology.

There is little work done in speaker recognition in any of the Indian languages. In this paper, we describe a text-dependent speaker recognition system for Telugu. Telugu is the second largest spoken language in the country.

2 Task Definition

The current application is in the context of a voice based personal messaging system. Initially, the user keys in a user-name by clicking on a specific sequence of icons on the touch screen. He/she then speaks out his/her password. The system is expected to recognize the user and display his/her photograph for double checking before continuing. Clearly, this scenario does not call for a high performance biometric for strict access control. Preventing the unauthorized access through stealing the password or mimicking the voice is less important than correctly identifying valid users. In case the recognition fails, it is merely a case of "wrong number" - there is no big risk.

The default user and his/her stored password are already known from the user-name. The first task, therefore, is to recognize the spoken password and verify it. An isolated-word speech recognition system is used for this purpose. Thereafter, a text-dependent speaker verification system is used to further verify the identity of the speaker within the given set of users in the system.

3 A Brief Survey of Speaker Recognition Research

One of the motivating factors behind research in speaker recognition is to understand how we human beings are able to recognize people by their voice. It is believed that the voice print of individuals is unique. Speaker recognition is closely related to other aspects of speech processing such as speech recognition, synthesis, coding, compression, etc.. These are inter-disciplinary areas borrowing from electrical engineering, digital signal processing, mathematics, computer science, linguistics, psychology and biology. Here we give a brief sketch of some of the relevant works.

Li Lui [8] shows that among the various parameters such as pitch, LPCC, ΔLPCC, MFCC, ΔMFCC that are extracted from speech signals, LPCC and MFCC are effective representations of a speaker. ΔLPCC and ΔMFCC are transitional spectral informations which alone are not suitable for speaker recognition whereas they can be used jointly with LPCC and MFCC respectively. Pitch can also be used in conjunction with other spectral features to improve the performance of speaker recognition.

Prosody is another feature that depends, to some extent, on the speaker. Michael J Carey [3] has exploited robust prosodic features for speaker identification. The performance of an existing HMM-based speaker recognition system could be increased by incorporating prosody, pitch and energy contours into the system.

Bing Xinag [13] proposes a method for speaker recognition which uses Gaussianization. Short-time Gaussianization is initiated by a global linear transformation of the features, followed by short-time windowed cumulative distribution function (CDF) matching. Linear transformation in feature space leads to decorrelation. CDF matching is applied to segments of speech localized in time and the aim is to warp the given feature so that its CDF matches normal distribution. Nowadays, speaker recognition systems based on Gaussian mixture models (GMM) are considered as highly robust and reliable.

Ran D Jilka [6] claims that text-independent speaker verification using covariance modelling is a viable alternative to GMM systems. This technique suggests two verification methods, namely, frame level scoring and utterance level scoring. Covariance of the features is calculated and then compared with that of reference model at frame level and at utterance level respectively. These methods have been shown to give better performance compared to GMM based system.

K Yu [15] compares Hidden Markov Models, Dynamic Time Warping and Vector Quantizations for speaker recognition. For text-independent speaker recognition, VQ performs better then HMMs and for text-dependent speaker recognition, DTW outperforms VQ and HMM based methods. Increasing train data did not change the balance significantly.

Michael Inman [5] proposed a technique in which segment boundary information is derived from HMMs which in turn provides a means of normalizing the formant patterns. Phonetic tempo variability (the tendency of the constituent phonetic segments of a word to vary in length from one occasion of speaking to another) and variability over time (tendency of the speech of a speaker to change as a function of time, typically days) have been addressed using cochlear filters and HMMs.

Chi Wei Che [4] proposed a HMM based text-prompted speaker verification system. In this method each speaker has a separate set of HMMs for each phoneme and the system uses concatenated phoneme HMMs. It has been shown that three-state single mixture phone HMMs produce better performance than single state tie-mixture Gaussian models. Another approach to speaker verification using HMMs was proposed by Michael Savic [11], which was based on adaptive vocal tract model which emulates the vocal tract of the speaker.

A New set of features named Adaptive Component Weighting (ACW) cepstral coefficients are introduced by Khaled T Assaleh [1]. These features emphasize the formant structure of the speech spectrum while attenuating the broad-bandwidth spectral components. ACW spectrum introduces zeros into the usual all-pole linear predictive spectrum. This is same as introducing a Finite Impulse Response (FIR) filter that normalizes the narrow band modes of the spectrum. ACW features have been evaluated on text-independent speaker identification system and shown to yield good performance.

Vector Quantization based Gaussian modelling and, training VQ codebook for HMM-based speaker identification have also been proposed to improve the performance of existing systems [9, 7].

The problem of speaker recognition for Indian languages has not been explored much. Here we describe a system we have designed and built for speaker recognition using Telugu data.

4 A Two Stage Password Authentication System

The current task involves password recognition using an isolated word speech recognition system and then further verification of the speaker knowing the word spoken. These two stages are described in order.

Speech technologies are generally language specific. Language to language variations in utterance of a given word can be quite marked. Fixing a language would enable language models to be employed. In our experiments here we have used Telugu data throughout. Needless to say, the proposed system is itself not tied down to Telugu or any particular language.

4.1 Recognizing the Password

A password file is created, containing usernames and corresponding passwords. Each user is asked to choose his password and is requested to record his password 10 times. A speaker-independent isolated-word speech recognition system is built with the combined data of all the users.

After the usual pre-processing steps [10], Mel Frequency Cepstral Coefficients are calculated using *HCopy* command of HTK. The process of calculating the features of the speech samples is called coding of the data. A HMM model is then created for each monophone from a prototype model with required model topology with means 0 and variances 1. Global means and variances are calculated from the coded data using the *HCompV* command. In the process of pruning, the values of the means and variances are updated using the *HERest* command of HTK. Once the monophone models are created and the parameters re-estimated sufficient number of times, triphone models are created and the triphone transcriptions are obtained running the *HLEd* command on the monophone transcriptions.

Due to insufficient data associated with many states, the variances in the output distributions might have been floored. So, within the triphone sets, the states need to be tied in order to share data and be able to make robust parameter estimates [14]. A decision tree based state tying is done by *HHEd* command of HTK. This process is called cloning of the models. The triphone models are further re-estimated over several iterations to create more robust models.

M Sukumar [12] explains the procedure to build a speaker-independent continuous speech recognition system for Telugu using the *HTK* tool kit. Only the changes required to adapt this standard procedure for our task are given below.

Grammar

The structure of the grammar file should be

$$\textit{\$word} = pw_1 \mid pw_2 \mid \ \ldots \ \mid pw_n; \\ \textit{(SENT\_START \$word SENT\_END)}$$

where $pw_1, pw_2, \ldots, pw_n$ are the passwords of n speakers. This structure ensures that the recognizer outputs only one password which best matches the given speech signal.

Dictionary

As our system is set to recognize isolated-words, we do not make an entry for 'sp' in the dictionary. (While running *HDMan* command of HTK make sure that *global.ded* script file is not present in the current directory, because, this appends 'sp' at the end of every pronunciation in the dictionary.) There will be an entry for each password in the dictionary.

As the pronunciations of all the words are present in the dictionary, the word level transcriptions of all the speech files are created and the phone level transcriptions are obtained by invoking *HLEd* command of HTK. All these transcription files must follow the standard HTK format of *mlf* files.

The isolated-word speech recognition system thus built, is used to recognize the password spoken.

4.2 Verifying the Speaker

Once the password is verified, we need to check whether the person who spoke the password is the right person or not.

An isolated-word speech recognition system is built separately for each speaker. Let us name these systems $S_1, S_2, \ldots, S_n$. For each system, the training data includes only the recordings of that particular speaker. The word list file and the grammar file of each system contains only one word which is the password of that speaker. Hence the same word is recognized whatever be the input. The degree of match indicated by the recognizer is used to verify the speaker's identity.

For the purpose of recognition, *HVite* command of HTK is used. This outputs, along with the recognized word, average log probability per frame a and total log probability for each test sample t. HTK automatically marks the silence at the beginning and ending of the speech signal and the total log probability is calculated for the utterance. Average log probability per frame and the total log probability are chosen because they together indicate the speaker, the password, and his/her speaking rate.

For each system S_i, with all the speech samples of the corresponding speaker $s_1, s_2, \ldots, s_m$ the pattern matrix

$$P_{2\times m} = \begin{bmatrix} s_1\_a & s_2\_a & \ldots & s_m\_a \\ s_1\_t & s_2\_t & \ldots & s_m\_t \end{bmatrix}$$

is computed. The mean $\mu = (\mu_a, \mu_t)$ and covariance matrix are calculated, where

$$\mu_a = \frac{1}{m}\sum_{j=1}^{m} a_j$$

and

$$\mu_t = \frac{1}{m}\sum_{j=1}^{m} t_j$$

A covariance matrix is a symmetric matrix which shows the correlation between the elements of the vectors. In our case, the covariance matrix for the set of two dimensional vectors is

$$\begin{bmatrix} \sigma_{aa} & \sigma_{at} \\ \sigma_{ta} & \sigma_{tt} \end{bmatrix}$$

For any $p \in \{a,t\}$ and $q \in \{a,t\}$, σ_{pq} represents how element p of the vector changes with respect to q. The value of each element of the covariance matrix is calculated by the formula given below.

$$\sigma_{pq} = \sum_{m}\{(p_m - \mu_p) \times (q_m - \mu_q)\}$$

Let us define a speaker model as M_i as $M_i = \{S_i,\ \mu_i,\ \Sigma_i\}$.

Mahalanobis distance is used to find the similarity between a test sample and a speaker model. This distance is based on correlations between data items, by which different patterns can be identified and analyzed. It is a way of determining similarity of an unknown sample to a known one. It differs from Euclidean distance in that it takes the correlations of the data items into account. Formally, the Mahalanobis distance from a group of values with mean $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ and covariance matrix Σ for a multivariate vector $x = (x_1, x_2, \ldots, x_n)$ is defined as,

$$d(x) = \sqrt{(x-\mu)^t \Sigma^{-1} (x-\mu)}$$

Euclidean distance weighs each component of the vectors equally, whereas Mahalanobis distance measure standardizes the data items so that differences in scale between the data items do not affect the distances. If the covariance matrix is the identity matrix then Mahalanobis distance is same as Euclidean distance.

The distance d_i from (a_i, t_i) to the group of values with mean μ_i and covariance matrix Σ_i is computed for $i = 1, 2, \ldots, n$. Let

$$K = \arg\min_i \{ \ (d_i) \ \}$$

Then K is declared as the identity of the speaker for a given test data.

5 Experiments and Results

In our experiments here, we have used data from 7 speakers. Each speaker is asked to choose any Telugu word as his password and is requested to record that word 20 times. 10 samples are used for training and the remaining 10 samples are used for testing. To check the performance of the speaker recognition system in the case of many users having the same password, the Telugu word "*SubhAkAnkshalu*", a kind of greeting, is also recorded 20 times by each of the speakers.

Initially the **isolated-word speech recognition system** is tested for password recognition. This system is built with speech samples of all the speakers with different passwords.

number of speakers = 7
number of training samples = 7 × 10 = 70
number of test samples = 70

The output of the *HResults* command of HTK for this test data is shown below.

```
======= HTK Results Analysis
======
Date: Tue Aug 16 04:49:45 2005
Ref : tstref.mlf
Rec : recout.mlf
——————— Overall Results ———————
SENT: %Correct=98.57 [H=69, S=1, N=70]
WORD: %Corr=98.57, Acc=98.57 [H=69,
D=0, S=1, I=0, N=70]
==========================
```

Here, the sentence level accuracy and word level accuracy are same because it is an isolated-word speech recognizer. The results show that out of 70 test samples **98.57%** samples are correctly recognized.

To evaluate the performance of the **speaker recognition system**, we need to check for 2 cases.

Case 1: Different passwords for different speakers:

For each speaker i, we randomly took 10 samples of his password recordings as training data and built the system S_i and a model M_i. The remaining 10 samples of each speaker are given for testing. 10 fold cross validation is performed. Performance is also measured on training data. The percentage accuracy for each case is noted in Table 1.

Table 1. Different Passwords for Different Speakers

Test no.	% Accuracy for Training Samples	% Accuracy for Test Samples
1	100.0 %	98.57 %
2	100.0 %	97.14 %
3	97.14 %	85.71 %
4	100.0 %	94.28 %
5	100.0 %	94.28 %
6	98.57 %	92.85 %
7	100.0 %	98.57 %
8	98.57 %	84.28 %
9	98.57 %	87.14 %
10	100.0 %	100.0 %
Avg. =	99.28 %	**93.71 %**

It may be observed that the average performance of the system is **93.71%**

Case 2: Same password for all speakers:

The performance of the system when different users use the same password is depicted in Table 2.

It may be noted that the average performance of the system in this case is **91.85%**. Although the performance has slightly come down, this shows that the system is able to differentiate between different speakers even if they all speak the same password.

Table 2. Same Password for All Speakers

Test no.	% Accuracy for Training Samples	% Accuracy for Test Samples
1	100.0 %	91.42 %
2	100.0 %	100.0 %
3	100.0 %	84.28 %
4	100.0 %	100.0 %
5	100.0 %	95.71 %
6	98.57 %	85.71 %
7	98.57 %	84.28 %
8	98.57 %	95.71 %
9	98.57 %	91.42 %
10	100.0 %	90.00 %
Avg. =	99.42 %	**91.85 %**

The performances of the two stages of the password authentication system were given individually above only to show the strengths of the two stages separately. In actual usage, however, the second stage is attempted only if the test sample is approved in the first stage.

6 Conclusions

In this paper we have proposed a method for text-dependent speaker recognition. This method uses HMM based modelling for each speaker. MFCC coefficients have been used as features. The HTK tool kit is used in the implementation. Mahalanobis distance has been employed.

A two stage password authentication system is built to illustrate the working of the speaker recognition system. In the first stage the password spoken is recognized and checked whether it matches with the stored password or not. In the second stage the speaker who spoke the word is recognized and verified.

Results obtained are promising.

There is little work done in speaker recognition in any of the Indian languages. Here we have described a text-dependent speaker recognition system for Telugu. Good performance has been achieved. Although Telugu data has been used to build the current system, the same procedure can be followed to build the system for any language. Preliminary experiments suggest that incorporating a Speech end-point detector can further improve the performance of the system.

References

[1] Khaled T Assaleh. New lp-derived features for speaker identification. In *IEEE Transactions on Speech and Audio Processing*, volume 2, pages 630–638, 1994.

[2] R Auckenthaler. Language dependency in text-independent speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings of ICASSP '01*, volume 1, pages 441–444, 2001.

[3] Michael J Carey. Robust prosodic features for speaker identification. In *Fourth International Conference on Spoken Language. Proceedings of ICSLP '96*, volume 3, pages 1800–1803, 1996.

[4] Chi Wei Che. An hmm approach to text-prompted speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings of ICASSP '96*, volume 2, pages 673–676, 1996.

[5] Michael Inman. Speaker identification using hidden markov models. In *Fourth International Conference on Signal Processing, Proceedings of ICSP '98*, volume 1, pages 609–612, 1998.

[6] Ran D Jilka. Text-independent speaker verification using covariance modeling. In *IEEE Signal Processing Letters*, volume 8, pages 97–99, 2001.

[7] ZHANG Lingua. A new method to train vq codebook for hmm-based speaker identification. In *7th International Conference on Signal Processing. Proceedings of ICSP '04*, volume 1, pages 651–654, 2004.

[8] Li Lui. Signal modeling for speaker identification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings of ICASSP '96*, volume 2, pages 665–668, 1996.

[9] J Pelecanos. Vector quantization based gaussian modelling for speaker verification. In *Proceedings of 15th International Conference on Pattern Recognition*, volume 3, pages 294–297, 2000.

[10] L Rabiner. *Fundamentals of Speech Recognition.* Prentice-Hall, 2003.

[11] Michael Savic. Variable parameter speaker verification system based on hidden markov modelling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings of ICASSP '90*, volume 1, pages 281–284, 1990.

[12] M Sukumar. Building a speaker independent continuous speech recognition system for telugu. Master's thesis, DCIS, University Of Hyderabad, 2005.

[13] Bing Xiang. Short-time gaussianization for robust speaker verification. In *IEEE International Conference on Acoustics,*

Speech, and Signal Processing. Proceedings of ICASSP ‘02, volume 1, pages 681–684, 2002.

[14] S Young. *The HTK Book (for HTK Version 3.2)*, 2002.

[15] K Yu. Speaker recognition using hiddem markov models, dynamic time warping and vector quantisation. In *IEEE Proceedings - Vision, Image, and Signal Processing*, volume 142, pages 313–318, 1995.