Text-Dependent Speaker Recognition System for Indian Languages

R. Rajeswara Rao[†] **A. Nagesh** **Kamakshi Prasad** **K. Ephraim Babu**

JNT University Hyderabad, India
JNT University Hyderabad, India
JNT University Hyderabad , India
University of Hyderabad Hyderabad, India

Summary

Speaker Recognition is a process of automatically recognising 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 holds great promise. Speech based recognition permits remote access. Speech is very important in a country like India with a large population with low levels of literacy and education. Speech empowers people by helping them to overcome the barriers of language and complexity of usage.

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 given set of users. Mel Frequency Cepstral Coefficients have been used to build Hidden Markov Models. The HTK tool kit has been used to build these systems. Mahalanobis distance measure is employed. Experiments conducted are described and the results are shown. Mostly spoken ten Indian languages speech data has been used in these experiments. It can be seen that the system gives very good performance for the intended task.

Key words:
Hidden Markov Models, Speaker Recognition, HTK.

1 Introduction

Speaker Recognition is a process of automatically recognising who is speaking on the basis of speaker dependent features of the speech signal. Speaker recognition system must be robust against mimicking. As of today, finger prints and retinal scans are more reliable means for identification. Nevertheless, during the years ahead, it is hoped that speaker recognition technology will make it possible to reliably verify the person's identity. Other methods such as finger prints and retinal scans cannot be used for some applications which run remotely such as telephone based automated control services. Speech is the only biometric that can be used for secure remote accessing.

Speaker Recognition is 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 categorised as *text-independent* and *text-dependent*. By text-independence, it is meant that the speaker can speak any utterance in a particular language. Whereas in the case of text-dependent systems the speaker is required to speak predefined piece of text such as a specific password.

The decision of the speaker verification system is just binary, whether to accept or not. If we represent the probability of the given utterance x belonging to the i^{th} speaker by $p_i(x)$, then the task of speaker verification is,

$$\text{if } p_i(x) > \textit{Threshold} \text{ then ACCEPT}$$
$$\text{else REJECT}$$

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

On the other hand, speaker identification system must decide among the N trained speakers. If $p_i(x)$ is the probability of the given utterance x belonging to i^{th} speaker, then 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 that speaker recognition systems can make. They are, *false acceptance* and *false rejection*. If a wrong person is verified or

Manuscript received November 5, 2007
Manuscript revised November 20, 2007

identified, then this is a false acceptance. Whereas, if a right person is rejected or not identified, then it is a false rejection. In the case of speaker verification, the decision of acceptance and rejection is made based on the threshold. So, the system can be designed to reduce 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 increase by this, unauthorized accessing 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.

Today, there are around one billion people in India, speaking 150 different languages. There is little work done in speaker recognition in Indian languages. In this paper, we describe a text-dependent speaker recognition system for mostly spoken ten Indian languages [2].

2 Task Definition

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.

The current application is in the context of a voice based personal messaging system. Initially, the user keys in a username 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 recognise the user and display his/her photograph for double checking before continuing. Clearly, this scenario does not call for a high performance biometric for tight 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" and nothing very dangerous happens. The default user and his/her password are already known from the username. The first task, therefore, is to recognise 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

The motivation behind the research in speaker recognition is to understand how we human beings are able to recognise people by their voice. 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.

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 information’s 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 depend, 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 localised in time and tries 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 modeling 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.

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 ocassion 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 coffcients 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 normalises the narrow band modes of the spectrum. ACW features have been evaluated on text-independent speaker identification system and shown to yeild 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.

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.

4.1 Recognizing the Password

The system is provided with a password file which has the entries of usernames and their corresponding passwords. Each user is asked to choose any word as his password and is requested to record his password 10 times. With this combined data of all the users an isolated-word speech recognition system is built. We have used mostly spoken ten Indian languages, namely Telugu, Hindi, Urdu, Kannada, Marathi, Tamil, Malayalam, Bengali, Oriya and English. Words as all of our research in speech technologies is currently focused based these Indian Languages. Needless to say, the proposed system is not tied down to any Indian Language or any particular language.

After the usual pre-processing steps [10], the Mel Frequency Cepstral Coeffcients of all the training samples 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 re-estimated sufficient number of times, triphone models are created using these monophone models and the triphone transcriptions obtained by executing *HLEd* command on the monophone transcriptions.

The triphone HMMs created so far share a common transition matrix. Due to the insufficient data associated

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

The following paragraphs explain the procedure to build text-dependant speaker recognition system for Indian language using the HTK *tool kit*.

Grammar

The structure of the grammar file would be

$$\$word = pw_1 \mid pw2 \mid \ldots\ldots. \mid pw_n;$$
$$(SENT\_\ START\ \ \$word\ \ SENT\_\ END)$$

Where $pw_1, p_{w2,} \ldots\ldots. pw_n$ are the passwords of n speakers. This structure ensures that the recogniser 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 transcritions 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 recognise the word spoken.

4.2 Verifying the Speaker

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

An isolated-word speech recognition system is built 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 recognised 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 remaining portion. These two values are chosen because they combinely represent the speaker, 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

$$P_{2xm} = \begin{bmatrix} s_{1\_a} & s_{2\_a} & \cdots & s_{m\_a} \\ s_{1\_t} & s_{2\_t} & \cdots & s_{m\_t} \end{bmatrix}$$

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

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

The 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

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

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.

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

Let us define a speaker model as M_i as

$M_i = \{ \mu_i, \sum_i \}$.

Mahalanobis distance is used to find the similarity measure between a test sample and speaker model. This distance is based on correlations between data items by which different patterns can be identified and analysed. It is a useful way of determining similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes the correlations of the data items into account. Formally, Mahalanobis distance can be defined as dissimilarity measure between two random vectors *x* and *y* of the same distribution with the covariance matrix $\sum$ is defined as,

$$d(x, y) = \sqrt{(x - y)^t \sum^{-1} (x - y)}$$

Euclidean distance weights 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 form (a_i, t_i) to the group of values with mean μi and covariance matrix $\sum_i$ is computed for i = 1, 2 … n.

Let

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

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

5 Experiments and Results

The Data Set that is created consists of the recordings of 10 speakers. Each speaker is asked to choose one word from a particular Indian language as his password and is requested to record that word for 20 times, so that 15 samples can be used for training and the remaining 5 samples are used for testing.

Initially the isolated-word speech recognition system that was built to recognize the password is tested. This system is built with speech samples of all the speakers with different passwords.

number of speakers = 10
number of training samples = 10 × 15 = 150
number of test samples = 10 × 5 = 50

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

```
======= HTK Results Analysis ==================
Date: Wed Oct 16 04:49:45 2007
Ref : tstref.mlf
Rec : recout.mlf

---------------------- Overall Results-------------------------
SENT: %Correct=98.57 [H=69, S=1, N=50]
WORD: %Corr=98.57, Acc=98.57 [H=69,
D=0, S=1, I=0, N=50]
========================================
```

Here, the sentence level accuracy and word level accuracy are same because it is an isolated-word speech recogniser. The results show that out of 150 test samples 98.57% samples are correctly recognised.

Evaluation

For each speaker i, we randomly took 15 samples of his password recordings as training data and built the system S_i and a model M_i. The remaining 5 samples of each speaker are given for testing. The behavior of the system with the training data is also noted by giving the training samples as test data. We repeated this process for motioned the ten Indian languages.

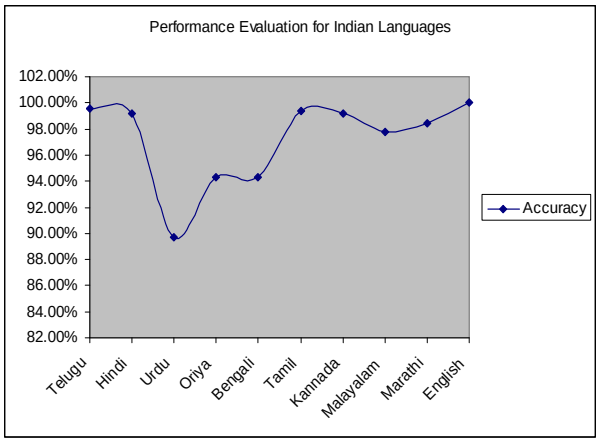
To check the consistency of the performance obtained, a 10 fold cross validation is performed. The percentage accuracy for each Indian language are noted in the form of tables.

Table1. Performances evaluation for Indian Languages

Language	% Accuracy for Training Samples	% Accuracy for Test Samples
Telugu	100.0%	99.57%
Hindi	100.0%	99.14%
Urdu	97.14%	89.71%
Oriya	99.18%	94.28%
Bengali	98.44%	94.28%
Tamil	100.0%	99.41%
Kannada	99.6%	99.23%

Malayalam	98.94%	97.8%
Marathi	99.87%	98.4%
English	100.0%	100.0%
Avg =	**99.32%**	**97.18%**

The total Performance of the system = 97.18%



6 Conclusions

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

A password authentication system is built to demonstrate the working of the speaker recognition system. This system works in two stages. In the first stage the word spoken is recognised and checked whether it matches with the password or not. In the second stage the speaker who spoke the word is recognized and verified. This system is built and tested for various training and test samples. The experiments that are conducted and their results are listed.

Most spoken ten Indian languages data has been used to build the system. The same procedure can be followed to build the system for any language. Speech endpoint detector can be incorporated to further improve the performance of the system.

Acknowledgments

Our sincere thanks to Mr. M Ravindranath for his valuable suggestions and cooperation.

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 veri_cation 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 veri_cation. 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.
[16] Douglas O'Shaughnessy, *Speech communications Human and Machine* Universities Press.2nd edition.

R.Rajeswara Rao received the B.Tech from Nagarjuna University. and M.Tech from JNT University degrees in Computer Science and in 1999 and 2003, respectively. He is currently pursuing Ph.D from JNT University, Hyderabad, India since 2004. His research areas of interest are Speech Processing, Neural Networks, and Pattern Recognition.

A.Nagesh received the B.Tech and M.Tech from Osmania University degrees in Computer Science and in 1996 and 2002, respectively. He is currently pursuing Ph.D from JNT University, Hyderabad, India since 2004. His research areas of interest are Genetic Algorithms, Speech Processing and Pattern Recognition.