International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 01, January 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

NEW FEATURE VECTORS FOR AUTOMATIC TEXT-INDEPENDENT SPEAKER TRACKING SYSTEM USING HIDDEN MARKOV MODELS

Dr A. Nagesh

Computer Science & Engineering,

Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad (India)

ABSTRACT

The objective of this paper to explore new feature vectors for Speaker tracking using LP analysis and Mel frequency Cepstral coefficients (MFCC). In this paper a new method is proposed to extract a set of source features for speaker tracking system using Hidden Markov Models (HMMs). LP analysis is used to extract the source information from the speech signal, which is speaker specific. From the speech signal speaker specific source characteristics are captured using Linear Prediction (LP) residual signal of speech signal. Next the MFCC features are extracted from source features. This new type of feature vectors contains the prosody and speaker specific information which are useful for speaker tracking. Using this new set of features experiments are carried out using HMMs for varying number of states of varying number of Gaussian mixtures.

Keywords: Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Linear Prediction (LP), Mel Frequency Cepstral Coefficients (MFCC), Speaker Tracking System (STM).

I. INTRODUCTION

Speaker tracking is an important issue in multimedia applications requiring analysis of spoken documents. In this paper, we propose a speaker tracking system (STS) which contains two main steps. First, a speaker segmentation process detects speaker change. This process works with no use of a priori knowledge about two speakers. Then, after segmentation, a classical speaker verification system is applied on each segment to determine whether it has been uttered by the target speaker.

Speaker tracking system aims to detect speech segments corresponding to known target speakers in a known audio resource. Three major applications of speaker tracking are broadcast news, meetings or seminars and telephone conversations. Speaker tracking system finds the conversations between several speech sources (persons) in which few are already enrolled to speech tracking system and other are unknown speech sources, and a target speaker will be chosen in a set of enrolled users. In the first step, speech is segmented in to knowledge about speakers. Then, the resulting segments are checked to belong to one of the target speakers.

In this paper, a speaker tracking and detection system for TIMIT dataset used for the target speaker. The conversation is modeled as a two speaker and hidden Markov model (HMM). Speaker tracking is used to segment speakers for detection, which is carried out by averaging frame scores of the Viterbi path.

II. FEATURE EXTRACTION

The feature vectors are represented in the form of Gaussians. Using GMM as a front end the feature vectors are extracted from the speech signal. For any system the basic requirement is to obtain the feature vectors from the speech signal. In the literature is found that some attempts are made to explore the new way of representing the feature vector based on the GMM feature extraction. The feature vectors are represented in probability vectors form. Instead representing GMMs as scalar value, it is represented as a probability vector. So the system performance is improved.

For Speaker tracking task, the new feature vectors are obtained from the speech signal estimating using probability density function based on Gaussian mixture model. The underlying speaker specific discrimination information is represented as a Gaussians.

Beginning from the training data of speaker S_i, a 12 dimensional feature vectors are extracted with a frame size of 25ms and frame shift of 10ms. These feature vectors are grouped into clusters with 'R' Gaussian mixtures as shown in Fig.1.

Fig. 1: R Gaussians for Speaker S_i.

2.1 Computation of New Feature Vectors

Once 'R' Gaussians, R clusters are formed. Each cluster represented as one Gaussian. The feature vector $X=(X_1, X_2,\ldots X_{12})$ is passed through a Gaussian G_1 by calculating probability P_1 using probability density function of Gaussian G_1. This P_1 is first coefficient in the new feature vector. In the same way feature vector X is passed through R Gaussians by creating R feature vector coefficients namely $P_1, P_2, \ldots, P_R$, as shown in Fig. 2. These R coefficients create a new feature vector of dimension R.

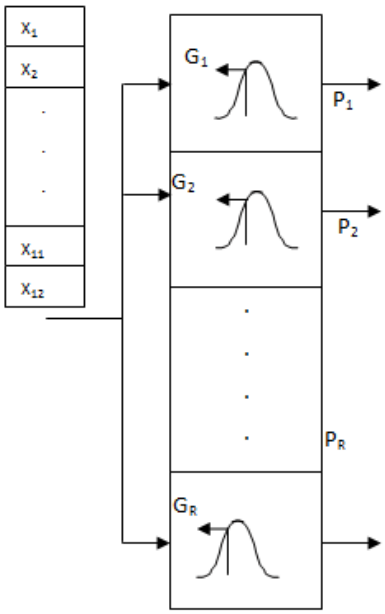

Fig.2: Parameter Estimation for New Feature Vector P. When R=20, The Good Identification Performance Has Been Achieved.

In this way, all the feature vectors are passed through ' R ' Gaussians (G_1, G_2,……., G_R) generating new R dimensional feature vectors. In other words the 12 dimensional MFCC feature vectors of size 'N' are transformed to 'R' dimensional feature vectors of size N. The 12 dimensional MFCC feature vector is represented as a 'R' dimensional feature vector.

In the new feature vector, each Gaussian probability density represents one coefficient. Experiments are carried out to find the dimension of new feature vector for good language identification performance. This is done by varying the number of Gaussians (coefficients) from 15 to 30, i.e number of coefficients in the new feature vector. When the number of coefficients are 12, the good recognition performance is achieved. The 12 dimensional MFCC feature vector is represented as a 12 dimensional new probability feature vector as shown in fig.3. The newly derived feature vectors are given to the HMM based classifier for speaker identification.

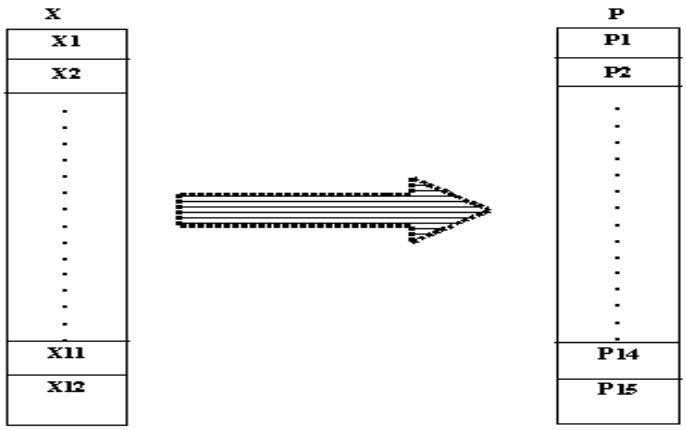

Fig.3: Transforming from 12 Dimensional MFCC feature Vector to 15 Dimensional Feature Vector

III. SPEAKER TRACKING USING HIDDEN MARKOV MODEL

Hidden Markov model (HMM) describes a two-stage stochastic process. The first stage consists of a Markov chain. In the second stage then for every point in time t an output or emission (observation symbol) is generated. This sequence of emissions is the only thing that can be observed of the behavior of the model. In contrast, the state sequence taken on during the generation of the data cannot be observed.

The Hidden Markov Models (HMM) is a doubly embedded stochastic process where the underlying stochastic process is not directly observable. The HMM not only models the underlying speech sounds, but also the temporal sequencing among the sound units. This temporal modeling is advantageous for speaker tracking task. The model parameters can be collectively represented as $\lambda_i = \{A_i, B_i, \pi_i\}$. The model parameters can be collectively represented as $\lambda_i = \{A_i, B_i, \pi_i\}$ for i = 1, … M. Each speaker in a speaker tracking system can be represented by a HMM and is referred to by the speaker respective models as λ_i.

In this HMM based speaker tracking system the type of HMM considered is Ergodic HMM. The main advantage of using ergodic HMM is it not only captures the positional patterns, but also temporal patterns effectively. Hence, to capture both categories of underlying patterns, continuous ergodic HMM is proposed for speaker tracking task.

At the beginning of HMM training, the HMM parameters $\lambda_i = \{A_i, B_i, \pi_i\}$ are initialized as follows. The states transition probability *A* and initial state observation probability are initialized randomly. The observation matrix *B* is initialized using *K*-means clustering algorithm. Using these initial values of the HMM parameters, further improvement is achieved with a Baum-welch re-estimation procedure. This way the new feature vectors of speaker S_i are trained to create one HMM for each speaker. This procedure is repeated for all the speaker under consideration and separate HMM is created for each speech signal.

During testing, the process of new feature vector generation is identical to training phase feature vector generation. In testing, scores are evaluated using HMM models. The $p(P|l)$ for each model is calculated, where $P = (P_1, P_2, \ldots, P_T)$ is the sequence of the test feature vectors. The language model that gives the highest score is declared as the identified language.

IV. EXPERIMENTAL EVALUATION

The speaker track system is the task of identifying who said what in the speech signal. In this work TIMIT speaker database is used. The TIMIT speech corpus contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the united states. For this study 100 female and 100 male are considered for the task. For training 45 sec of speech data and testing 1 sec of speech data is used.

4.1 Experimental Setup

The speaker tracking system has been implemented using Matlab7. In this paper we considered 3-state HMM for the identification task. For each state, the varying Gaussian components such as 2,4,8 and 16 are considered. For testing different test utterances such as 1 sec, 2 sec and 3 sec are taken. We considered LP order of 12 for all experiments. The speaker tracking system rate is defined as the ratio of the number of speakers identified to the

total number of speakers tested. The speaker recognition performance for 2 state, 3-state and 4-state performance increased uniformly. The test speech duration increased, the performance also increased.

Table1.Speaker Tracking System Performance

No. Of States	No. of Mixture Components	Recognition Rate (%)		
		1 Sec	2 Sec	3 Sec
2	2	88	89	91
	4	90	92	93
	8	91	92	94
3	2	90	92	93
	4	91	93	95
	8	93	94	96
4	2	93	93	94
	4	91	94	96
	8	94	95	96

V. CONCLUSION

In this paper we demonstrated the importance of new form of source information for speaker tracking system. Linear residual based on source information is represented as a probability vector. The performance of speaker track system effectively capture the speaker-specific in formation form LP residual. The experiments are carried out varying the test duration and number of Gaussians in each state in HMM. The performance this new feature vector based speaker tracking system is superior when compared to the conventional feature vector based speaker tracking system.

REFERENCES

[1] Lie Lu and Hong-Jiang Zhang, Speaker change detection and tracking in real-time news broadcasting analysis, in ACM International Conference on Multimedia, 2002, pp. 602–610.

[2] I-M. Chagnolleau, A-E. Rosenberg, and S.Parthasarathy, Detection of target speakers in audio databases, in ICASSP'99, 1999.

[3] D.A.Reynolds, The Effects of Handset Variability on Speaker Recognition Performance Experiments on the Switchboard Corpus, In Proc. Int. Conf. on Acoust., Speech and Signal Processing, vol.1(1996), pp. 113117 .

[4] B. Xiang, Text-independent speaker verification with dynamic trajectory model, IEEE Signal Processing Letters 10 (May 2003), 141 -143.

[5] Q. Jin, T. Schultz, and A. Waibel, Speaker identification using multilingual phone strings, In Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2002) , vol. 1(May 2002), pp. 145-148.

[6] M. Zissman, Comparison of four approaches to automatic language identification of telephone speech, IEEE Trans. on Speech and Audio Processing , vol. 1 (January 1996), pp. 31-44.

[7] K.N. Stevens, Acoustic Phonetics. Cambridge, England: The MIT Press, 1999.

[8] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of IEEE, vol. 77, no. 2, pp. 257-286, Feb.1989.