www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 8 August, 2014,Page No. 7874-7878

Robust Feature Based Automatic Text-Independent Gender Identification System Using Ergodic Hidden Markov Models(HMMs)

R. Rajeswara Rao*
*Department of CSE, JNTU Kakinada-Vizianagaram, AP, India
rajaraob4u@gmail.com

ABSTRACT

In this paper, robust feature for Automatic text-independent Gender Identification System has been explored. Through different experimental studies, it is demonstrated that the timing varying speech related information can be effectively captured using Hidden Markov Models (HMMs). The study on the effect of feature vector size for good Gender Identification demonstrates that, feature vector size in the range of 18-22 can capture Gender related information effectively for a speech signal sampled at 16 kHz, it is established that the proposed Gender Identification system requires significantly less amount of data during both during training as well as in testing. The Gender Identification study using robust features for different states and different mixtures components, training and test duration has been exploited. I demonstrate the Gender Identification studies on TIMIT database.

Keywords - Gaussian Mixture Model (GMM), Gender, LPC, MFCC.

INTRODUCTION

With the development of more and more identification systems to identity a Gender, there is a need for the development of a system which can provide identification task such as gender identification automatically without any human interface. Gender identification using voice of a person is comparatively easier than that from other approaches. There exist several algorithms for automatic gender identification but none of them has found to be 100% accurate. Gender Identification System can be represented like any other pattern recognition system as shown in Fig. 1. This task involves three phases, feature extraction phase, training phase and testing phase [1]. Training is the process of familiarizing the system with the voice characteristics of a speaker, whereas testing is the actual recognition task.

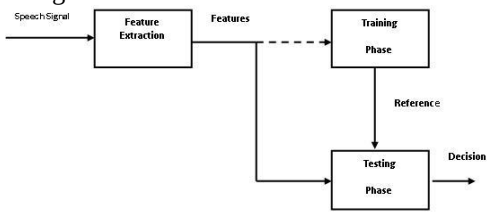

Fig. 1: A typical Block diagram representation of a Gender Identification task.

In Gender identification based on the voice of a speaker consists of detecting if a speech signal is uttered by a male or a female. Automatically detecting the gender of a speaker has several potential applications. In the context of Automatic Speech Recognition, gender dependent models are more accurate than gender independent ones [1] [2]. Hence, gender recognition is needed prior to the of speaker recognition. In the context of speaker recognition, gender detection can improve the performance by limiting the search space to speakers from the same gender. Also, in the context of content based multimedia indexing the speaker's gender is a cue used in the annotation. Therefore, automatic gender detection can be a tool in a content-based multimedia indexing system.

Much information can be inferred form a speech, such as sequences of words, gender, age, dialect, emotion, and even level of education, height or weight etc. Gender is an important characteristic of a speech. Automatically

detecting the gender of a speaker has several potential applications such as (1) sorting telephone calls by gender (e.g. for gender sensitive surveys), (2) as part of an automatic speech recognition system to enhance speaker adaptation, and (3) as part of automatic speaker recognition systems. In the past, many methods of gender classification have been proposed. For parameters selections, some methods used gender dependent features such as pitch and formants [3] [5].

Speech is composite signal which has information about the message, gender, the speaker identity and the language [6][7]. It is difficult to isolate the speaker specific features alone from the signal. The speaker characteristics present in the signal can be attributed to the anatomical and the behavioral aspects of the speech production mechanism. The representation of the behavioral characteristics is a difficult task, and usually requires large amount of data. Automatic speaker recognition systems rely mainly on features derived from the physiological characteristics of the speaker.

Speech is produced as sequence of sounds. Hence the state of vocal folds, shape and size of various articulators, change over time to reflect the sound being produced. To produce a particular sound the articulators have to be positioned in a particular way. When different speakers try to produce same sound, through their vocal tracts are positioned in a similar manner, the actual vocal tract shapers will be different due to differences in the anatomical structure of the vocal tract. System features represent the structure of vocal tract. The movements of vocal folds vary from one speaker to another. The manner and speed in which the vocal folds close also varies across speakers. Hence different voices are produced. Source features represent these variations in the vibrations of the vocal folds.

The theory of Linear Prediction (LP) is closely linked to modeling of the vocal tract system, and relies upon the fact that a particular speech sample may be predicted by a linear combination of previous samples. The number of previous samples used for prediction is known as the order of the prediction. The weights applied to each of the previous speech samples are known as Linear Prediction Coefficients (LPC). They are calculated so as to minimize the prediction error. As a byproduct of the LP analysis, reflection coefficients and log area coefficients are also obtained [8].

A study into the use of LPC for speaker recognition was carried out by Atal [9]. These coefficients are highly correlated, and the use of all prediction coefficients may not be necessary for speaker recognition task [10]. Sambur [11] used a method called orthogonal linear prediction. It is shown that only a small subset of the resulting orthogonal coefficients exhibits significant variation over the duration of an utterance. It is also shown that reflection coefficients are as good as the other feature sets. Naik et. al., [12] used principal spectral components derived from linear prediction coefficients for speaker verification task. Hence a detailed exploration to know the speaker-specific excitation information present in the residual of speech is needed and hence the motivation for the present work.

I. EXPLORING ROBUST FEATURES FOR GENDER IDENTIFICATION

Here, the GMM is used as front-end to extract features vectors from speech signal. For the Gender Identification ASR task, the basic requirement is to obtain the feature vectors form the speech signal. Recently, some attempts are made to explore the alternative representation of feature vectors based on GMM feature extraction.

For Speaker Recognition task, robust features are derived from the speech signal based on estimating a Gaussian mixture model. The underlying speaker discrimination information is represented by Gaussians. The estimated GMM parameters means, co-variance and component weight can be related to the formant locations, bandwidths and magnitudes.

For the proposed new feature vectors, from the speech signal of a speaker S_i, a 12 dimensional MFCC feature vectors are obtained with a window size of 20ms and window shift of 3 ms. These MFCC feature vectors are distributed into 'R' Gaussians mixtures as shown in Fig. 2.

Fig. 2: R Gaussians for Speaker S_i.

The feature vector X=(X1, X2,……, X12) is passed through a Gaussian G1 by calculating a Gaussian probability P1 using Gaussian probability density function. This P1 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 P1,P2,….,PR, as shown in Fig. 3. These R coefficients create a new R dimensional feature vector. The newly created R dimensional feature vector is shown in the Fig. 4.

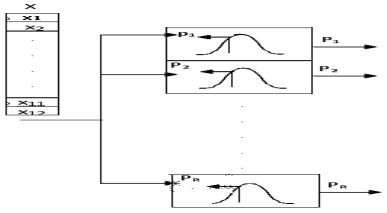

Fig. 3: Parameter estimation for new vector P.

When R=14, the optimal recognition performance has been achieved.

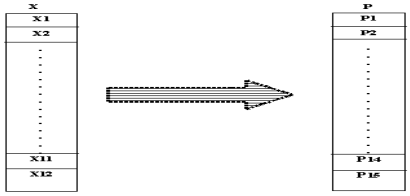

Fig. 4: Transforming from 12 dimensional MFCC feature vector to R dimensional feature vector.

Experiments are carried to find the dimension new feature vector for good speaker recognition performance. This is done by varying the number of Gaussians from 12 to 30, i.e. number of coefficients in the new feature vectors. When the numbers of coefficients are 20, the good identification performance is achieved [4].

II. CONTINUOUS ERGODIC HIDDEN MARKOV MODEL FOR SPEAKER RECOGNITION

The HMM is a doubly embedded stochastic process where the underlying stochastic process is not directly observable. HMMs have the capability of effectively modeling statistical variations in spectral features. In a variety of ways, HMMs can be used as probabilistic speaker models for both text-dependent and text-independent speaker recognition [17][18]. HMM not only models the underlying speech patterns but also the temporal sequencing among the sounds. This temporal modeling is advantageous for text-dependent speaker recognition system. Left Right HMM can model temporal sequence of patterns only, where as to capture the patterns of different type ergodic HMM is used [19]

As shown in the Fig. 4 in the training phase, one HMM for each speaker is obtained (i.e., parameters of model are estimated) using training feature vectors. The parameters of HMM are [MA, et.al, 2007] State-transition probability distribution: It is represented by $A = \lfloor a_{ij} \rfloor$

Where

$$a_{ij} = P(q_{t+1} = j \mid q_t = i) \quad 1 \le i, j \le N \qquad (2)$$

defines the probability of transition from state i to j at time t.

For a three state left-right model the state transition matrix is given as

$$A = \{a_{ij}\} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} \quad (3)$$

The state transition matrix of three state ergodic model is given by

$$A = \{a_{ij}\} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad (4)$$

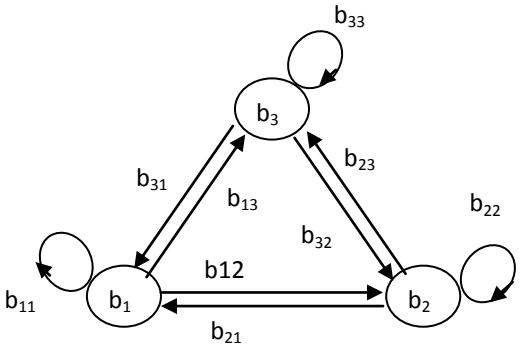

Fig. 5:Three-state ergodic HMM.

Observation symbol probability distribution: It is given by $B = \lfloor b_j(k) \rfloor$ in which

$$b_j(k) = P(O_t = V_k \mid q_t = j) \quad 1 \le k \le M \quad (5)$$

defines the symbol distribution in state $j = 1,2,3....N$. The initial state distribution is given by $\pi = P(q_1 = i)$ where

$$\pi_i = P(q_1 = i) \quad 1 \le i \le N \quad (6)$$

Here, N is the total number of states, and q_t is the state at time t, M is the number of distinct observation symbols per state, and O_t is the observation symbol at time t. In testing phase, $P(O/\lambda)$ for each model is calculated, where $O = (O_1 O_2 O_3 .... O_T)$ Here the goal is to find out the probability for a given model to which the test utterance belongs to. The speaker whose model gives the highest score is declared as the identified speaker. GMM corresponds to a single-state continuous ergodic HMM.

The model parameters can be collectively represented as $\lambda = (A_i, B_i, \pi_i)$ for $i = 1.....M$. Each speaker in a speaker identification system can be represented by a HMM and is referred to by the speaker's respective models λ.

In the testing phase, p (O/λ) for each model is calculated [21]. where O=(o1o2o3…OT) is the sequence of the test feature vectors. The goal is to find the probability, given the model, that the test utterance belongs to that particular model. The speaker model that gives the highest score is declared as the ident

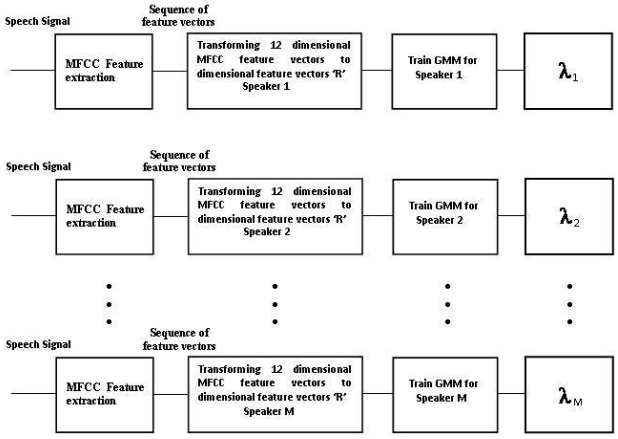

Fig. 6: Training HMM for Gender Recognition Task

estimation is not possible and therefore a special case of ML estimation known as Expectation-Maximization (EM) [K. N. Stevens, 1999] algorithm is used to extract the model parameters.

The GMM likelihood of a sequence of T training vectors $X = \{\bar{x}_1, ... \bar{x}_T\}$ can be given as [16]

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\bar{x}_t \mid \lambda)$$

The EM algorithm begins with an initial model λ and tends to estimate a new model $\overline{\lambda}$ such that $p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$ [21]. This is an iterative process where the new model is considered to be an initial model in the next iteration and the entire process is repeated until a certain convergence threshold is obtained

III. EXPERIMENTAL EVALUATION

A. Database used for the study

Gender identification is the task of identifying whether the speaker is male or female. In this paper we consider identification task for TIMIT Speaker database [16].

The TIMIT corpus of read speech has been designed to provide speaker data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speaker recognition systems. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. We consider 100 male speakers and 100 female out of 630 speakers for gender recognition. Maximum of 30 sec. of speech data is used for training and minimum of 1 sec. of data for testing. In all the cases the speech signal was sampled at 16 kHz sampling frequency. Throughout this study, closed set identification experiments are done to demonstrate the feasibility of capturing the Gender -discrimination information from the speech signal. Requirement of significantly less amount data for Gender-discrimination information and Gaussian mixture models is also demonstrated.

B. Experimental Setup

The system has been implemented in Matlab7 on Windows XP platform. We have trained the model GMM using Gaussian Components as 2, 4, 8, and 16 for training speech duration of 10, 20 and 30 sec. Testing is performed using different test speech durations such as 1 sec., 2 sec., and 3 sec..

II. Performance Evaluation

The system has been implemented in Matlab7 on windows XP platform. The result of the study has been presented in Table 1. We have used Vector order of 18 for all experiments. We have trained the model using Gaussian mixture components as 4, 8, 16, 32 and 64 for training speech lengths as 20 sec.,. Testing is performed using different test speech lengths such as 1 sec, 3 sec, and 5 sec.. Here, recognition rate is defined as the ratio of the number of genders identified to the total number of genders tested. As shown in Table. 1 the identification rate for testing length for 5 sec. outperformed, where as for testing length of 3 sec. is also on par with 5 sec. testing length. Table. 1, shows identification rate increases when different number of mixture components 4, 8, 16, 32 and 64 with different test speech lengths 1 sec., 3 sec., and 5 sec..

The percentage (%) recognition of Gaussian Components such as 4, 8, 16, 32 and 64 seems to be uniformly increasing. The minimum number of Gaussian components to achieve good recognition performance seems to be 32 and thereafter the recognition performance is minimal. The recognition performance of the HMM drastically increases for the test speech duration of 1 sec. to 3 sec.. Increasing the test speech duration from 3 sec. to 5 sec. improves the recognition performance with small improvement.

Table 1: Gender Recognition Performance for 20 Sec. Training speech duration

No. of States	No. of Mixture Components	Speaker Recognition (%) Test Duration (in sec.)		
		1 Sec.	3 Sec.	5 Sec.
2	4	74	88	94
	8	82	95	98
	16	84	96	98
	32	86	97	99.5
	64	84	94	97
3	4	96	98	98.5
	8	98	98.5	100
	16	98.5	100	100
	32	99	99.5	99
	64	97	98	98.5
4	4	95	96	98
	8	94	96.5	98
	16	96	98.5	99
	32	97	98	99
	64	95	97	99

IV. CONCLUSION

In this work we have demonstrated the importance of coefficient order for speaker recognition task. gender discrimination information is effectively captured for coefficient order 18 using a HMM .The recognition performance depends on the training speech length selected for training to capture the gender-discrimination information. Larger the training length, the better is the performance, although smaller number reduces computational complexity.

The objective in this paper was mainly to demonstrate the significance of the gender-discrimination information present in the speech . We have not made any attempt to optimize the parameters of the model used for feature extraction, and also the decision making stage. Therefore the performance of speaker recognition may be improved by optimizing the various design parameters.

REFERENCES

[1] Alex Acero and Xuedong Huang, "*Speaker and Gender Normalization for Continuous-Density Hidden Markov Models"*, in Proc. of the Int. Conf. on Acoustics, Speech, and Signal , IEEE, May 1996.

[2] C. Neti and Salim Roukos., "*Phone-specific gender-dependent models for continuous speech recognition"*, Automatic Speech Recognition and Understanding Workshop (ASRU97), Santa Barbara, CA, 1997.

[3] R. Vergin, A. Farhat and D.O'Shaughnessy, "*Robust gender-dependent acoustic-phonetic modeling in continuous speech recognition based on a new automatic male/female classification*", Proc. Of IEEE Int. Conf. on Spoken Language (ICSLP), pp. 1081, Oct. 1996.

[4] A. Nagesh, V. Kamakshi Prasad, " New Feature Vectors for Automatic Text-Independet Language Identification" IJCSC, Vol. 2, Issue 2 pp553-558, 2011.

[5] S. Slomka and S. Sridharan, "*Automatic gender identification optimized for language independence*", Proc. Of IEEE TENCON'97, pp. 145-148,Dec. 1997.

[6] O'Shaughnessy D., 1987.,"*Speech Communication: Human and Machine",* Addison-Wesley, New York.

[7] Rabiner L.R., Juang B.H., 1993. "*Fundamentals of Speech Recognition",* Prentice-Hall, Englewood Cliffs, NJ.

[8] Makhoul, J., 1975. "*Linear prediction: a tutorial review",* Proc. IEEE 63, 561–580.

[9] B.S. Atal, "*Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification*", J. Acoust. Soc. Ameri., *vol. 55*, pp.1304-1312, Jun. 1974.K. Elissa, "Title of paper if known," unpublished.

[10] A.E. Rosenberg and M. Sambur, "*New techniques for automatic speaker verification*", *vol. 23, no.2*, pp.169-175, 1975.

[11] M. R. Sambur, "*Speaker recognition using orthogonal linear prediction*", IEEE Trans. Acoust. Speech, Signal Processing, *vol. 24*, pp.283-289, Aug. 1976

[12] J. Naik and G. R. Doddington, " *High performance speaker verification using principal spectral components*", in proc. IEEE Int. Conf. Acoust. Speech, Singal Processing, pp. 881-884, 1986.

[13] Gish H., Krasner M., Russell W., and Wolf J., "*Methods and experiments for text-independent speaker recognition over telephone channels*", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), *vol. 11*, pp. 865-868, Apr. 1986.

[14] Reynolds D. A., and Rose R. C., " *Robust Text-Independent Speaker Identification using Gaussian Mixture Models*'', IEEE-Transactions on Speech and Audio Processing, *vol. 3, no. 1*, pp. 72-83,1995.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "*Maximum likelihood from incomplete data via the EM algorithm*", J. Royal Statist. Soc. Ser. B. (methodological), *vol. 39*, pp. 1-38, 1977

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

[17] M. Forsyth, "Discriminating observation probability (dop) HMM for speaker verification", Speech Comm., vol. 17, pp. 117-129, 1995.

[18] R, Rajeshwara Rao, "Automatic Text -Independent Speaker Recognition using source features", Ph.D. thesis., Jan-2010.

[19] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition. Prentice-Hall, 1993.

[20] MA, B., Li, H., and Tong, R. Spoken language recognition with ensemble classifier. IEEE Trans. Audio, Speech and Language Processing 15, 7 (September 2007), 2053-2062.)

[21] A. P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. Royal Statist. Soc. Ser. B. (methodologies), vol. 39, pp. 1-38, 1977.

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

`