Source Feature Based Gender Identification System Using GMM

R. RAJESHWARA RAO **#1**, NAGESH ***2,**
DVR College of Engineering & Technology
Department of CSE, Hyderabad, AP, India
*MG Institute of Technology
Department of CSE, Hyderabad, AP, India

Abstract— In this paper, through different experimental studies it is demonstrated that the excitation component of speech can be exploited for text independent gender identification system. Linear prediction (LP) residual is used as a representation of excitation information in speech. The speaker-specific information in the excitation of voiced speech is captured using Gaussian Mixture Model (GMM). The decrease in the error during training and recognizing correct gender during testing demonstrates that the excitation component of speech contains speaker-specific information and is indeed being captured by GMM. It is demonstrated that the proposed gender identification system using excitation information requires significantly less amount of data both during training as well as in testing, compared to the other gender identification systems. A gender identification study using source feature for different Mixtures Components, train and test duration has been exploited. We demonstrate the gender identification studies on TIMIT database.

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

I. INTRODUCTION

With the development of more and more identification systems to identity a person, there is a need for the development of a system which can provide personal 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.

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, gender dependent model. 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] [4].

Speech is composite signal which has information about the message, gender, the speaker identity and the language [5][6]. 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 behavioural aspects of the speech production mechanism. The representation of the behavioural 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 modelling 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 [7].

A study into the use of LPC for speaker recognition was carried out by Atal [8]. These coefficients are highly correlated, and the use of all prediction coefficients may not be necessary for speaker recognition task [9]. Sambur [10] 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., [11] 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.

The rest of the paper is organized as follows: In Section II we examine the gender characteristics of the LP residual, and discuss issues involved in extracting the speaker-specific information from the residual. In Section III we discuss feature extraction using Mel Ceptral coefficients to capture the speaker specific information from the residual. Section IV describes Gaussian Mixture Model for Gender Recognition. Section V describes the database used in the study and Section V1 describes performance evaluation of Gender identification sytem. The proposed gender recognition system, based on the LP residual, may not require large amounts of data.

II. GENDER CHARACTERISTICS IN THE LP RESIDUAL

Speech signals, as any other real world signals, are produced by exciting a system with source. A simple block diagram representation of the speech production mechanism is shown in the Fig.1. Vibrations of the vocal folds, powered by air coming from the lungs during exhalation, are the sound source for speech. Hence, as can be from Fig. 1, the glottal excitation forms the source, and the vocal tract forms the system. One of the most powerful speech analysis technique is the method of linear predictive analysis. The philosophy of linear prediction is intimately related to the basic speech production model. The Linear Predictive Coding (LPC) analysis approach performs spectral analysis on short segments of speech with an all-pole modelling constraint [12]. Since speech can be modelled as the output of linear, time-varying system excited by a source, LPC analysis captures the vocal tract system information in terms of coefficients of the filter representing the vocal tract mechanism. Hence, analysis of speech signal by LP results in two components, namely the synthesis filter on one hand and the residual on the other hand. In brief, the LP residual signal is generated as a by product of the LPC analysis, and the computation of the residual signal is given below.

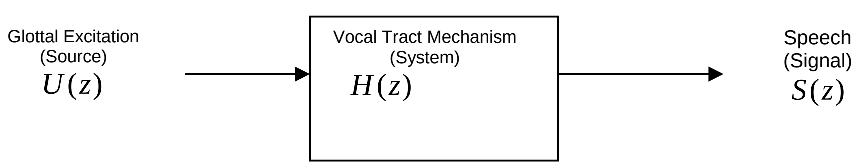

Fig. 1: Source and System representation of speech production mechanism.

If the input signal is represented by u_n and the output signal by s_n, then the transfer of the system can be expressed as,

$$H(z)=\frac{S(z)}{U(z)} \qquad (1)$$

Where $s(z)$ and $u(z)$ are z-transforms of s_n and u_n respectively.

Consider the case where we have output signal and the system and have to compute the input signal. The above equation can be expressed as $S(z) = H(z)U(z)$

$$U(z) = \frac{S(z)}{H(z)} \qquad (2)$$

$$U(z) = \frac{1}{H(z)} S(z) \qquad (3)$$

$$U(z) = A(z)S(z) \qquad (4)$$

Where $A(z) = {}^{1}\!/_{H(z)}$ is the inverse filter representation of the vocal tract system.

Linear prediction models the output s_n as the linear function of past outputs and present and past inputs. Since prediction is done by a linear function, the name linear prediction. Assuming an all-pole for the vocal tract, the signal s_n can be expressed as linear combination of past values and some input u_n as shown below.

$$Sn = -\sum_{k=1}^{p} a_k S_{n-k} + GU_n \qquad (5)$$

Where G is a gain factor.

Now assuming that the input u_n is unknown, the signal s_n can be predicted only approximately from a linear weighted sum of past samples. Let this approximation of s_n be , where

$$\tilde{S}_n = -\sum_{k=1}^{p} a_k s_{n-k} \qquad (7)$$

Then the error between the actual value Sn and predicted value is given by $e_n = S_n - \tilde{S}_n$ [13]. This error is nothing but LP residual of signal is shown in Fig 2.

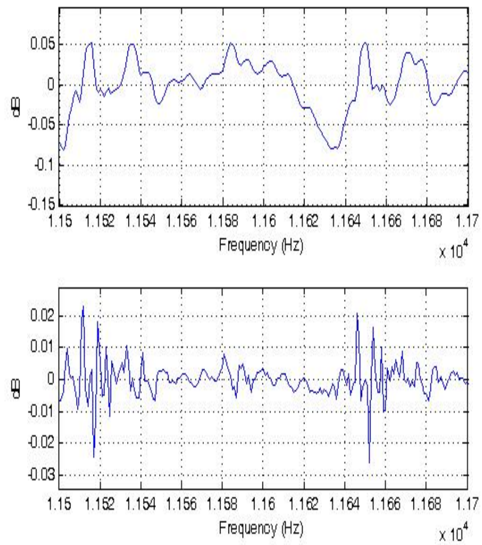

Fig 2: Actual signal and its LP residual

III. FEATURE EXTRACTION OF LP RESIDUAL SIGNAL

MFCC is the best known and most popular, and this features has been used for gender identification. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. Rather than the speech waveforms themselves, MFFCs are less susceptible to the said variations [6].

A. Motivation to use Melfrequency Cepstral coefficients(MFCC)

Since our interest is in capturing global features which correspond to source excitation, the low frequency or pitch components are to be emphasized. To fulfil this requirement it is felt that MFCC are most suitable as they emphasize low frequency and de-emphasize high frequencies.

B. MFCC

In this phase the digital speech signal is partitioning into segments (frames) with fixed length 10-30 ms from which the features are extracted due to their spectral qualities. Spectrum is achieved with fast Fourier transformation [14]. Then an arrangement of frequency range to mel scale follows according to relation

$$f_{mel} = 2595\log\left(1+\frac{f_{Hz}}{700}\right) \qquad (9)$$

By logarithm of amplitude of mel spectrum and applying reverse Fourier transformation we achieve frame cepstrum:

$$mel-cepstrum(frame) = FFT^{-1}\left[mel(\log|FFT(frame)|)\right]$$

The FFT-base cepstral coefficients are computed by taking IFFT of the log magnitude spectrum of the Speech signal. The mel-warped cepstrum is obtained by inserting a intermediate step of transforming the frequency scale to place less emphasis on higher frequencies before taking the IFFT [7][15][16].

IV. GAUSSIAN MIXTURE MODEL FOR GENDER IDENTIFICATION

GMM is a classic parametric method best used to model gender identities due to the fact that Gaussian components have the capability of representing gender information effectively. Gaussian classifier has been successfully employed in several text-independent gender identification applications. As shown in Fig. 3 in a GMM model, the probability distribution of the observed data takes the form given by the following equation [17][18].

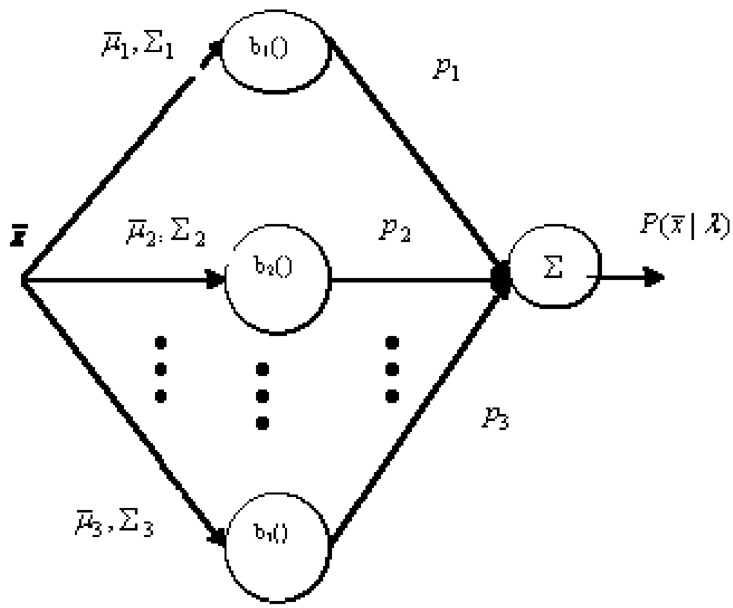

Fig. 3: Gaussian Mixture Model

$$p(\bar{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\bar{x}) \qquad (10)$$

Where M is the number of component densities, $\bar{x}$ is a D dimensional observed data (random vector), $b_i(\bar{x})$ are the component densities and p_i are the mixture weights for i = 1, .., M.

$$b_i(\bar{x}) = \frac{1}{(2\pi)^{D/2} \left| \sum_i \right|^{1/2}} \exp\left\{ -\frac{1}{2} (\bar{x} - \bar{\mu}_i)^T \sum_i {}^{-1} (\bar{x} - \bar{\mu}_i) \right\} \qquad (11)$$

Each component density $b_i(\bar{x})$ denotes a D-dimensional normal distribution with mean vector $\bar{\mu}_i$ and covariance matrix $\sum_i$. The mixture weights satisfy the condition $\sum_{i=1}^{M} p_i = 1$ and therefore represent positive scalar values. These parameters can be collectively represented as $\lambda = \{ p_i, \bar{\mu}_i, \Sigma_i \}$ for i = 1 … M. Each language in a language system can be represented by a GMM and is referred by the language respective model λ.

The parameters of a GMM model can be estimated using maximum likelihood (ML) [19] estimation. The main objective of the ML estimation is to derive the optimum model parameters that can maximize the likelihood of GMM. Unfortunately direct maximization using ML estimation is not possible and therefore a special case of ML estimation known as Expectation-Maximization (EM) [19] algorithm is used to extract the model parameters.

The GMM likelihood of a sequence of T training vectors $X = \{\bar{x}_1, \ldots \bar{x}_T\}$ can be given as [19]

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\bar{x}_t \mid \lambda) \qquad (12)$$

The EM algorithm begins with an initial model λ and tends to estimate a new model $\bar{\lambda}$ such that $p(X \mid \bar{\lambda}) \geq p(X \mid \lambda)$ [19]. 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

V. 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 [20].

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. Through out this study, closed set identification experiments are done to demonstrate the feasibility of capturing the speaker-specific information from the speech signal. Requirement of significantly less amount data for the gender recognition using speaker-specific excitation 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..

VI. PERFORMANCE EVALAUATION

The system has been implemented in Matlab7 on windowsXP platform. The result of the study has been presented in Table 1. We have used LP order of 12 for all experiments. We have trained the model using Gaussian mixture components as 2, 4, 8, and 16 for different training speech lengths as 10 sec., 20 sec., and 30 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 Fig. 4 the recognition 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. Fig. 5 shows identification rate increases when different train speech length varies from 10 sec., 20 sec., 30 sec., and number of mixture components 2, 4, 8, and 16 with different test speech lengths 1 sec., 3 sec., and 5 sec..

The percentage (%) recognition of Gaussian Components such as 2, 4, 8, and 16 seems to be uniformly increasing. The minimum number of Gaussian components to achieve good recognition performance seems to be 8 and thereafter the recognition performance is minimal. The recognition performance of the GMM drastically increases for the test speech duration of 1 sec. to 2 sec.. Increasing the test speech duration from 2 sec. to 3 sec. improves the recognition performance with small improvement.

VII. CONCLUSION

In this work we have demonstrated the importance of information in the excitation component of speech (pitch) for gender recognition task. Linear prediction residual is used to represent the excitation information. Performance of the recognition experiments shows that Gaussian Mixture models can capture speaker-specific excitation information from the LP residual. Performance of the system for different Mixture components shows that the optimal mixture components are 8 for speech signals sampled at 16 kHz. The recognition performance depends on the training speech length selected for training to capture the speaker-specific excitation 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 speaker-specific excitation information (pitch) present in the linear prediction residual for speaker recognition. 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] S. Slomka and S. Sridharan, "Automatic gender identification optimized for language independence", Proc. Of IEEE TENCON'97, pp. 145-148,Dec. 1997.

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

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

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

[8] 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

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

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

[11] 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.

[12] Furui, S., 1997. Recent advances in speaker recognition. Pattern Recognition Lett. 18, 859–872.

[13] S.R.Mahadeva Prassana, Cheedella S. Gupta, B. Yegnanarayana. Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Communications Vol.48 (2006) pp.1243-1261.

[14] Dempster, A., Laird, N., and Rubin, D., "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, vol. 39, pp. 1-38, 1977.

[15] Molau, S., Pitz, M., Schluter, R., and Ney, H., "Computing Mel-frequency cepstral coefficients on the power spectrum," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 73-76, May. 2001.

[16] Picone, J. W., "Signal modeling techniques in speech recognition," Proceedings of IEEE, vol. 81, no. 9, pp. 1215-1247, Sep. 1993.

[17] 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.

[18] 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.

[19] 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

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

TABLE 1:GENDER RECOGNITION PERFORMANCE WITH TIMIT DATABASE
(MALE -100 + FEMALE-100)

Training speech duration (sec)	No. of mixture components	Recognition rate (%) Testing speech length		
		1 sec	3 sec	5 sec
10	2	89	91	92
	4	91	93	93.5
	8	92	93.5	93
	16	92.5	93.5	93
20	2	92	93	94
	4	93	93.5	94.5
	8	94	94	95
	16	95	94	95
30	2	93	94	96
	4	94	94.5	96.5
	8	94.5	95	97
	16	95	96	98

Fig. 4: Gender Recognition Performance for varying Mixtures

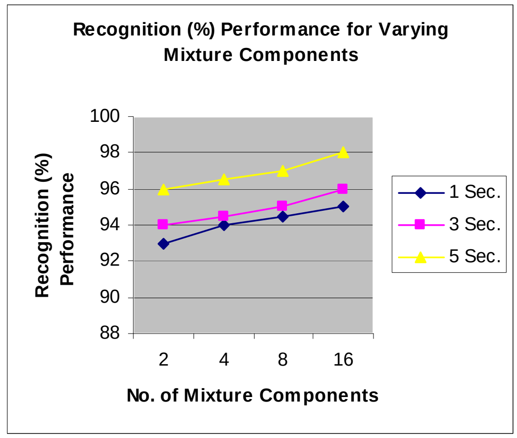

Fig. 5: Gender Recognition Performance for varying Train Duration