Gender Identification of a Speaker Using MFCC and GMM

Ergün Yücesoy[1], Vasif V.Nabiyev[2]

[1]Ordu Üniversitesi TBMYO, Ordu, Türkiye
yucesoye@hotmail.com
[2]KTÜ Bilgisayar Mühendisliği, Trabzon, Türkiye
vasif@ktu.edu.tr

Abstract

The gender of a speaker, which is the most distinctive characteristics of a speech, can easily be recognized by a person who hears it. Automatic identification of gender information from speech signal is substantially important for many applications. With the identification of gender, gender-dependent systems are defined and accuracy and robustness of systems can be increased. In this study, a system identifying the gender of a speaker independent from a text is developed. The proposed system is based on the classification of MFCC coefficients obtained from speech signals with GMM. In the study, the effect of Gaussian mixture number and MFCC coefficients to the system success is investigated. In the experiments by using TIMIT database, for the 760 sentences of 76 speakers 100 percent success, and for the 6100 sentences of 610 speakers 97.76 percent success is achieved.

1. Introduction

A listener, along with the content of a speech, can understand several characteristics of a speech such as his/her age, gender, accent and emotional state. The most specific characteristics in a speech is the gender of a speaker. If the gender of a speaker can automatically be identified, this information can be used in many fields. In the automatic speech recognition, the gender pre-information of a speaker makes it possible to define gender dependent models. Gender dependent models give more successful results than gender independent models [1, 2].

Like in all recognition systems, also for gender identification speech is firstly converted to some measurement values representing the speech that are called as features. Selection of features that will be used in letter stages is the main factor among the others affecting the system success. Today, to represent speech signal various features are used. Most important features are energy, pitch frequency, format frequency [3,4], linear prediction coefficients (LPC), linear prediction cepstral coefficients (LPCC), Mel-Frequency cepstral coefficients (MFCC) and their derivatives [5,6]. Speech signal converted to features vectors are modeled by using various classification methods. Neural Networks (NN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), and Support Vector Machine (SVM) are the most commonly used classification methods in speech recognition [7,8,9].

Pitch frequency or fundamental frequency, which is the vibration rate of vocal cords, is one of the mostly used features for the identification of the gender of a speaker [10]. Pitch frequency is obtained from only voiced portions of speech and it is quietly dependent of voice quality. In addition, there is an intersection region for the pitch frequency of male and female speakers. Therefore pitch frequency is not enough alone for gender identification. So far, studies using various features and classification methods have been carried out. In a study, by means of pitch frequency and the first two formants, 88.42% success is achieved [3]. In the following studies, error rate is substantially reduced by using MFCC features along with pitch frequency [5], Performance of the features of pitch (F_0) and cepstral (LPCC MFCC and PLP) in gender identification are compared [11] and finally the success of MFCC feature vector dimension and the selected phonemes in gender recognition is analyzed for 10 Hindi vowels and 5 nasals. Its success range is obtained between 94.12 and 100 percent [12].

In this study, an automatic system identifying the gender of a speaker without depending the text is proposed. And the study, MFCC, GMM and TIMIT are used as features vector, classifier and speaker database respectively. Details of the methods used and the results are shown throughout the study.

2. Feature Extraction

Speech signal does not contain speech information only. At the same time, it contains information like age, gender, and emotional state that are related to the speaker [13]. At the first stage of all identification system, features are determined. At this stage speech signal is converted into measurement values that have distinctive characteristics and less variability representing speech characteristics. There are various methods for this conversion and they are classified as parametric and non-parametric models.

In parametric approach a speech production model is defined based on human speech production mechanism. LPC analysis of speech is generally used for this aim. LPC models speech as the series connection of an excitation source E(Z) and a spectral shaping filter H(Z) representing vocal tract. Excitation source and vocal tract filter are used as a feature by extracting them from speech signal by means of cepstral analysis. In non-parametric approach, Mel-Frequency cepstral coefficients (MFCC) which is based on human voice perception mechanism are used.

2.1. Mel-frequency cepstral coefficients

MFCC is one of the most frequently used features both in speech and speaker recognition [13, 14]. Stevens and Volkman (1940) experimentally showed that human hearing system perceives the frequencies linearly up to 1 KHz and logarithmically above it. Relationship between perceived frequency which is called Mel and actual frequency is given in as,

$$Mel(f) = 2595 * log\left(1 + \frac{f}{700}\right) \quad (1)$$

MFCC is a cepstral method which converts speech into parameters according to mel-scale. The method whose block diagram is given in Fig.1 has the following steps.

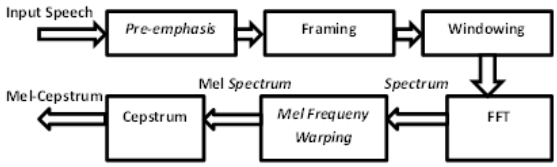

Fig. 1. MFCC block diagram

1. ***Pre-emphasis:*** Due to the structure of voice production system, damping occurs in high-frequency regions. For that reason, the spectrums of voiced regions are compensated by pre-emphasis which amplifies high-frequency regions and performs filtering [15]. Widely used pre-emphasis filter is given as,

$$Y[n] = x[n] - a * x[n-1], a \approx (0.95 - 0.97) \qquad (2)$$

In this study we took a=0.97

2. ***Framing And Windowing:*** Like in all voice analysis methods, also MFCC method is applied along the short portions where voice has stationary acoustic features [13]. These portions are generally selected as 20-30 milliseconds a shift of 10-15 milliseconds along the signal. Thus every frame contains a some portion of its previous frame. In voice applications Hamming window is generally preferred. Hamming window is expressed as,

$$w(n) = 0.54 - 0.46 * cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \qquad (3)$$

3. ***Frequency Spectrum****:* Speech signals divided into analysis windows are transformed by FFT from time domain to frequency domain. This notation representing frequency distribution of speech signal is called amplitude spectrum.

4. ***Mel-Frequency Warping:*** To convert the obtained amplitude spectrum into mel-scale, a filter set placed linearly with respect to mel-scale is used. This set consists of triangle band pass filters that are overlapping 50% as shown in Fig.2. Generally, filter coefficient is selected between 20 and 30.

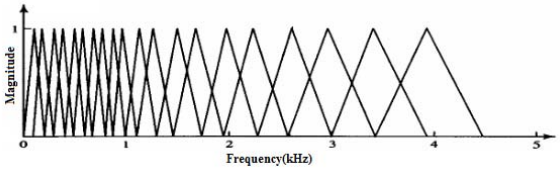

Fig. 2. Mel filter bank

5. ***Mel Spectrum And Cepstrum:*** At this stage, mel spectrum is obtained by multiplying amplitude spectrum of the signal with mel filter set and taking the logarithm of energy that is present in each filter. Since the logarithms of mel-spectrum coefficients are real numbers, discrete cosine transform given in equation (4) is used to go back to time domain. As a result, the obtained coefficients are called as mel-frequency cepstrum coefficients (MFCC).

$$\widetilde{c_n} = \sum_{k=1}^{K} (log\, \widetilde{S_k}) cos\,(n(k - \tfrac{1}{2})\tfrac{\pi}{K}), n = 1,2,\dots,K \qquad (4)$$

Here $\widetilde{S_k}$, k=1,2,...,K are mel spectrum coefficients. Since the first component obtained as a result of transformation represents the average logarithmic energy, is extracted from feature vector.

3. Gaussian Mixture Model

Although Gaussian distribution has some important analytic properties, for the modeling of real data sets substantial difficulties are encountered. If data set has two clumps only one Gaussian distribution cannot capture this structure but a linear combination of two Gaussian distributions characterizes this data better. By using a sufficient number of Gaussian, and by adjusting their means and covariances as well as coefficients in the linear combination, almost any continuous density can be approximated to arbitrary accuracy [16].

Mixture density created by weighted linear combination of K unimodal Gaussian components is expressed as,

$$p(\vec{x}|\lambda) = \sum_{k=1}^{K} p_k b_k(\vec{x}) \qquad (5)$$

Where $\vec{x}$ is a vector with D dimension, $b_k(\vec{x}), k = 1, \dots K$ are component densities and $p_k, k = 1, \dots K$ are mixture weights satisfying $\sum_{k=1}^{K} p_k = 1$. Density of each component is a Gaussian function with D dimension having the mean vector $\vec{\mu}_k$ and covariance matrix $\sum_k$ in equation (6).

$$b_k(\vec{x}) = \frac{1}{(2\pi)^{D/2}\sqrt{|\sum_k|}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_k)^T \sum_k{}^{-1} (\vec{x} - \vec{\mu}_k)\right\} \qquad (6)$$

GMM model is parametrically represented by depicting mean vector, covariance matrix and mixture weights of all components.

$$\lambda = \{p_k, \vec{\mu}_k, \sum_k\} \qquad (7)$$

3.1. Parameter Estimation

There are various methods for the estimation of the most suitable GMM parameters to the distribution of feature vectors obtained from the speech [17]. The most popular and well-established one of these methods is the Maximum likelihood (ML) estimation. The aim of ML estimation is to find the parameters that are maximizing the GMM likelihood from the given training data.

For a training data vector set of $X = \{\vec{x}_1, \dots, \vec{x}_T\}$ GMM likelihood can be written as,

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda) \qquad (8)$$

This expression is a nonlinear function of λ parameter and its direct maximization is impossible. For that reason, ML parameter estimation is obtained by an iterative way called as expectation maximization (EM) [18].

3.1. EM Algorithm

EM algorithm is a general method that is used for the ML parameter estimation from data sets where data are not complete or some data are missing. There are two basic applications of EM algorithm. The first one occurs in cases in which there really become some missing values. Missing values may originate from the problem of observation process or restrictions. The second one arises in case that analytic solution of likelihood function is difficult but it can be simplified by assuming that missing or hidden parameters exist. The latter one is more widely used in pattern recognition.

Let X be an incomplete data and assume that there is a random variable z= (z_1,...,z_k) with K-dimensional representing the identity of mixture components which produces X. In this case the mixture model is expressed as,

$$p(\vec{x}|\lambda) = \sum_{k=1}^{K} \alpha_k \, p_k(\vec{x}|z_k, \lambda_k) \quad (9)$$

where
$p_k(\vec{x}|z_k, \lambda_k)$ are the component densities defined by λ_k each,
z_k is an indicator variable having only single 1 and the rest 0,
And $\alpha_k = p(z_k)$ are the mixture weights whose sum ($\sum_{k=1}^{K} \alpha_k$) is 1. It represents the probability of any randomly selected $\vec{x}$ vector produced by k[th] component.

EM algorithm starts with the initial estimate of λ and continues by updating λ until the convergence is provided. Initial parameters and weights can be chosen by random or some heuristic methods. Each iteration consists of E and M steps.

E-Step: By using the current λ values, the probability w_{ik} of every data point $\vec{x}_i$ $1 \le i \le N$ belonging to $1 \le k \le K$ mixture component is calculated by the equation (10). For every data point $\vec{x}_i$, the sum of mixture weights is equal to $\sum_{k=1}^{K} w_{ik} = 1$. At this stage a NxK mixture weight matrix occurs whose each row summation is 1.

$$w_{ik} = p(z_{ik} = 1|\vec{x}_i, \lambda) = \frac{p_k(\vec{x}_i|z_k, \lambda_k)\alpha_k}{\sum_{m=1}^{K} p_m(\vec{x}_i|z_m, \lambda_m)\alpha_m} \quad (10)$$

M-Step: To compute new parameter values, firstly sum of mixture weights $N_k = \sum_{k=1}^{K} w_{ik}$ is computed for each component. This value is the effective number of data points that are assigned to k[th] component.

New mixture weight, mean and covariance matrix can be computed by the equations (11), (12) and (13) respectively.

$$\alpha_k{}^{new} = \frac{N_k}{N}, \; 1 \le k \le K \quad (11)$$

$$\overrightarrow{\mu_k}^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \cdot \vec{x}_i, \; 1 \le k \le K \quad (12)$$

$$\sum_k{}^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \cdot (\vec{x}_i - \overrightarrow{\mu_k}^{new})(\vec{x}_i - \overrightarrow{\mu_k}^{new})^T \quad (13)$$

After all the new parameters have been computed, again step E is returned and by computing member weights updating process for parameters continues. A pair of E and M steps is considered as one iteration. At the end of each iteration, logarithmic likelihood value which is defined by equation (14) is computed and if there becomes no substantial change at the end of iteration, convergence gets satisfied.

$$\log l(\lambda) = \sum_{i=1}^{N} log \, p(\vec{x}_i|\lambda) = \sum_{i=1}^{N} \left(log \sum_{k=1}^{K} \alpha_k \, p_k(\vec{x}_i|z_k, \lambda_k) \right) \quad (14)$$

Where $p_k(\vec{x}_i|z_k, \lambda_k)$ is the Gaussian distribution for the k[th] mixture component.

4. Experimentation and Results

In this study a system which determines the gender of the speaker automatically according to the speech signal is proposed. The system composed of two stages. In the first stage the system is trained by using the sentences that are vocalized by known gender speakers. In the second stage the system is tested by using the sentences that are vocalized by unknown gender speakers. The speech records used in this study are taken from TIMIT database. TIMIT database is composed of 10 sentences (2 are same and the others are different) that are spoken by 630 speakers (438 male, 192 female) who are speaking 8 different major dialects of American English. Some of the speakers of database are used for the training of the system and the rest is used for testing.

In the training stage by using the features vectors derived from the speech records determined for the training, GMM models are formed for male and female speakers each. GMM represent any distribution by the sum of weighted M Gaussian components. MFCC coefficients are used as features vector in this study. MFCC coefficients are calculated by speech signals (which are framed to 256 samples and 50% are overlapping) passing by 29 triangle filters that are placed linearly in mel-scale.

In the training stage, after forming the male and female models we pass to the test stage. In the test stage, MFCC features vectors derived from the test speakers are compared with the gender models of training stage. As a result of the comparison, a log-likelihood criterion which shows how the test data is fitted to the model is calculated. At the end, the gender of the speaker is determined by the model representing the larger log-likelihood value.

In the first experiment, the effects of the number of mixture and features dimensions are analyzed. For this reason, the trained system by using one sentences of 10 male and 10 female speakers, is tested by using 10 sentences of 76 speakers each. The results derived according to the number of mixture used and MFCC dimension are listed in table1.

Table 1. The effect of number of mixture and features dimention to the success

MFCC #	Mixture #	Correct #	Incorrect #	Success rate %
6	4	709	51	93,28
	8	718	42	94,47
	16	716	44	94,21
12	4	751	9	98,81
	8	757	3	99,6
	16	753	7	99,08
24	4	759	1	99,86
	8	758	2	99,74
	16	760	0	100

From the results of Table1, it is seen that the success rate increased from 93.4% with 6 MFCC coefficient and 4 Gaussian

components to 100% with 24 MFCC coefficient and 16 Gaussian components. By taking the computational load and success rate into consideration, the model of 12 MFCC coefficient represented by 8 Gaussian components is ideal for the proposed system.

The system is tested by 610 speakers who are not among the 20 speakers used for training in the second experiment that is made by using model parameters of the first experiment. 97.67% success is gained by correct determining of 5958 speech from 6100 speech of 610 speakers. It is understood that there is not an important change in the success rates both of the experiments and the system is not effected from the number of the speakers.

5. Conclusions

In this study the gender of the speaker is determined by modeling the MFCC coefficient derived from the vocalized sentence of the speaker with GMM. It is seen that the increase of the Gaussian components and MFCC coefficients are also increasing the success of the system. By analyzing the data amount used in the training of the system and the effects of the age distribution of the speakers, it is evaluated that the reliability and accuracy of the system will be increased.

6. 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. 1081-1084, 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] E.S. Parris and M.J. Carey, Language Independent Gender Identification, ICASSP, pp 685-688, 1996.

[6] Ting, H, Yingchun, Zhaohui, W., Combinning MFCC and Pitch to Enhance the Performance of the Geder Recognition, IEEE, 2006.

[7] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech and Audio Process., 3 (1), 72–83,Jan. 1995.

[8] M. H. Sedaaghi, "A Comparative Study of Gender and Age Classification in Speech Signals", Iranian Journal of Electrical & Electronic Engineering, Vol. 5, No. 1, pp. 1-12, March 2009.

[9] Djemili, Rafik, Hocine Bourouba, and Mohamed Cherif Amara Korba. "A speech signal based gender identification system using four classifiers." Multimedia Computing and Systems (ICMCS), 2012 International Conference on. IEEE, 2012.

[10] Abdulla, W. and Kasabov, N. 2001 . Improving speech recognition performance through gender separation. In Proc. Int. Conf. Artificial Neural Networks and Expert Systems (ANNES), pages 218–222, Dunedin, New Zealand.

[11] Pronobis, Marianna, and Mathew Magimai Doss. "Analysis of F0 and Cepstral Features for Robust Automatic Gender Recognition."

[12] Deiv, D. Shakina, and Mahua Bhattacharya. "Automatic Gender Identification for Hindi Speech Recognition.". International Journal of Computer Applications (0975 – 8887). Volume 31– No.5, October 2011

[13] L. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition, Englewood Cliffs (N.J.), Prentice Hall Signal Processing Series, 1993.

[14] J. R. Deller, J. H. L. Hansen, J. G. Proakis, Discrete-Time Processing of Speech Signals, IEEE Press, Piscataway (N.J.), 2000.

[15] Picone, J., Signal modeling techniques in speech recognition, Proceedings of the IEEE 81, pp. 1215–1247, 1993.

[16] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)

[17] G.McLachlan, Mixture Modeles. New York: Marcel Dekker, 1988

[18] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." Journal of the Royal Statistical Society. Series B (Methodological) (1977): 1-38.