Advances in Computational Research
ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, 2012, pp.-69-73.
Available online at http://www.bioinfo.in/contents.php?id=33

GENDER IDENTIFICATION USING SVM WITH COMBINATION OF MFCC

SANTOSH GAIKWAD, BHARTI GAWALI* AND MEHROTRA S.C.

Department of Computer Science & Information Technology, Dr.Babasaheb Ambedkar Marathwada University, Aurangabad, MS, India.
*Corresponding Author: Email- bharti_rokade@yahoo.co.in.

Received: February 21, 2012; Accepted: March 06, 2012

Abstract- Gender is an important and most diffrentiative characteristic of a speech. Gender information can also be used to improve the performance of speech and speaker recognition systems. Automatic gender classification is a technique that aims to determine the sex of the speaker through speech signal analysis. However with the increase in biometric security application, practical application of gender identification increased the many fold .The need of gender identification from speech arises several situation such as sorting telephonic call. Many methods of gender identification have been proposed in literature. We implemented the gender classification method and gender dependant feature such as pitch, roll of and energy in combination with MFCC. The clustered approach of above said parameter is implemented using SVM. We also present the experimental result of the proposed approach .It is observed that the accuracy of gender identification system is improved on the basis of size of codebook .The high accuracy is got at 25 codebook size with greater time slice. The accuracy of system tested with respective to gender and age .The efficient recognition rate of 95% is achieved in the age group of 25-30.
Keywords- Gender Identification, Pitch, Energy, MFCC, SVM

Citation: Santosh Gaikwad, Bharti Gawali and Mehrotra S.C. (2012) Gender Identification Using SVM with Combination of MFCC. Advances in Computational Research, ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, pp.-69-73.

Introduction

Gender identification based on the voice of a speaker consists of detecting a speech signal 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. Hence, gender recognition is needed prior to the application of one 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 annotation. Therefore, automatic gender detection can be a tool in a content-based multimedia indexing system. This paper describes an approach for voice-based gender identification for audio-visual content-based indexing. Several acoustic conditions exist in audio-visual data such as compressed speech, telephone quality speech, noisy speech, speech over background music, studio quality speech, different languages, and so on. Gender identification system must be able to process this variety of speech conditions with acceptable performance.

Gender identification is an important step in speaker and speech recognition systems [1-4]. In these systems, the gender identification step transforms the gender independent problem into a gender dependent one, thus it can reduce the size and complexity of the problem. [5, 6, 8, 9].

For speech signal based on gender identification, the most commonly used features are pitch period and Mel-Frequency Cepstral Coefficients (MFCC) [10]. The main intuition for using the pitch period comes from the fact that the average fundamental frequency (reciprocal of pitch period) for men is typically in the range of 100-146 Hz, whereas for women it is 188-221 Hz [11]. However, there are several challenges while using pitch period as the feature for gender identification. First, a good estimate of the pitch

period can only be obtained from voiced portions of a clean non-noisy signal [12, 13]. Second an overlap of pitch values between male and female.

For the problem of gender identification. Pitch estimation relies considerably on the speech quality. This drawback makes such an approach non-suitable for the problem of video indexing. Also, the reported results are based on with the signal of five second (5s) files which is not an image of the frame-based classification accuracy in a continuous speech signal.[1] Followed a general audio classifier approach using MFCC features and Gaussian Mixture Models (GMM) as a classifier. When applied to gender identification, the results are 73% of classification accuracy which is not promising. [4] Used a combination of pitch-based approach and general audio classifier approach using GMM. The reported results are based on 7s files after silence removal.

Previous studies on automatic gender classification from speech signals of adult speakers achieved high accuracy by using only features related to the fundamental frequency (F0) and the first four formant frequencies [5]. MFCC extracts the spectral components of the signal at 10ms rate by fast Fourier transform and carries out the further filtering based on the perceptually motivated Mel scale. In [14], the authors identified the gender of the speaker by evaluating the distance of MFCC feature vectors and reported identification accuracy of about 98%. However, using MFCC also has several limitations. First, MFCC captures linguistic Information such as words or phonemes at a very short timescale (several ms), increasing the computation complexity. Second, since MFCC learns too much detail about the short-time spectrum of the speech signal, it faces the problem of over-training; hence the performance of MFCC is significantly affected by recording conditions (like noise, microphone, etc.). For example, if the speech samples used for training and testing are recorded in different environments or with different microphones (a typical scenario in real world problems), MFCC fails to produce accurate results. To address the drawbacks of the above two approaches, techniques were proposed that combine both the pitch period and MFCC features discussed in [15], [16], [17]. However, the intrinsic drawbacks of the two features still affect the accuracy and computational complexity of the gender identification system.

In this paper, we propose a gender identification system that uses basic speech feature extracted from MFCC and gender dependant feature: pitch as a parameter selection. We estimated parameter classification using SVM.

The rest of the paper is organized as follows. In section database collection, we present the database collection. We addressed parameter extraction using MFCC and gender identification system in section parameter extraction using mfcc. We discuss parameter selection in parameter selection and described SVM classifier svm model section. The paper concludes with Experimental result and performance of system, conclusion with discussion.

Database Collection

The speech database collected from students of Department of CS & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad 20 speakers in which 8 were male and 12 were female. Each word in the vocabulary was recorded 5 times so that it will be good for training. In the vocabulary we selected as a real-time isolated word as well as natural continuous sentences.

Parameter Extraction Using MFCC

We are characterizing the signal in terms of the parameters of such a model, we must separate source and the model (filter). In ASR the source (fundamental frequency and details of glottal pulse) are not important for distinguishing different phones [18,19]. Instead, the most useful information for phone detection is the filter, i.e. the exact position and shape of the vocal tract. If we knew the shape of the vocal tract, we would know which phone was being produce to separate the source and filter (vocal tract parameters) efficient mathematical way is cepstrum. The cestrum is defined as the inverse DFT of the log. [20] The cepstral property have been extremely useful where the variances of different coefficients are tends to be uncorrected. The cepstral coefficients have the extremely useful property that variance of the different coefficients tends to be uncorrelated [21]. This is not true for the spectrum, where spectral coefficients at different frequency bands are correlated. The fact that cepstral features are uncorrelated means that the Gaussian acoustic model doesn't have to represent the covariance between all the MFCC features, which hugely reduces the number of parameters [22].The process of MFCC parameter extraction is explained in the equation 1 where x[n] is any input signal with limit value N= 0, 1…….N-1 , we got a log propagation C[n] Vector set presented in equation given below:

$$c[n] = \sum_{n=0}^{N-1} log\left(\left|\sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn}\right|\right) e^{j\frac{2\pi}{N}kn}$$

Where c(n) is cepstral coefficient and x(n) is the input signal. Since the MFCC is the most popular feature extraction technique for ASR [18], the basic steps involved in extraction of MFCC is shown in figure 1.

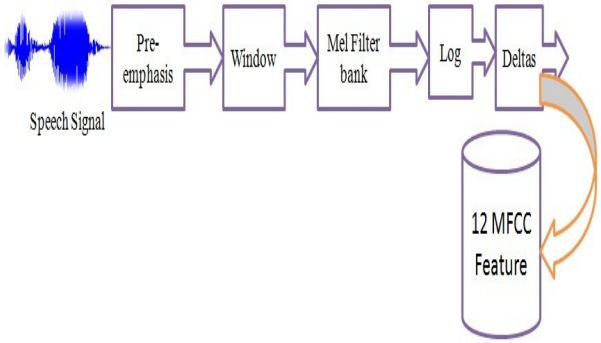

Fig 1- Steps for extracting a sequence of 12 MFCC feature vectors from waveform.

Parameter Selection

In the parameter selection we selected a basic MFCC 12 feature in addition to gender dependant pitch feature and basic supportive energy feature.

Energy of speech signal

The energy of speech is a basic and independent parameter, energy of each frame is calculated by equation given below

$$E_t = \int_{t}^{t+\tau} |X(t)|\, dt$$

Energy of all the frames is ordered and the top ones are selected for the following process to obtain the pitch feature. [23] The voiced frame and the sonorant frame were determined by calculating the energy contained within certain bandwidths. In our system, we just simply calculate the energy by using method described in (4).The computation complexity is greatly reduced. The following experimental results indicate that such a simple energy calculation is able to yield speech frames which contain relatively strong pitch feature.

Pitch Analysis

Pitch is defined as the fundamental frequency of the excitation source. Hence an efficient pitch extractor and an accurate pitch estimate calculated can be used in an algorithm of gender identification. The human voice is a magical tool. It can be used to identify those we know to create wonderful music through singing; it allows people to communicate verbally; and, it can help in the recognition of emotions. Everyone has a distinct voice, different from all others unique and can act as an identifier. The human voice is composed of a multitude of different components, making each voice different; namely, pitch, tone, and rate.

Support Vector Machine (SVM) Models

A Support Vector Machine (SVM) performs classification by constructing an N-dimensional hyper plane that optimally separates the data into two categories. SVM models are closely related to neural networks. In fact, a SVM model using a sigmoid kernel function is equivalent to a two-layer, perceptions neural network.

Support Vector Machine (SVM) models are a close cousin to classical multilayer perception neural networks. Using a kernel function, SVM's are an alternative training method for polynomial, radial basis function and multi-layer perception classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training.

In the parlance of SVM literature, a predictor variable is called an attribute, and a transformed attribute that is used to define the hyper plane is called a feature. The task of choosing the most suitable representation is known as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. So the goal of SVM modeling is to find the optimal hyper plane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyper plane are the support vectors.

Experimental Results

Experiments are carried out to validate the performance of the gender identification system proposed in this paper.

For the basic parameter extraction MFCC is used, with the following statics

No. of Coefficient: 12
Window Length=0.15
Time step:-0.5

For the annotation and normalization we used praat software as a tool.

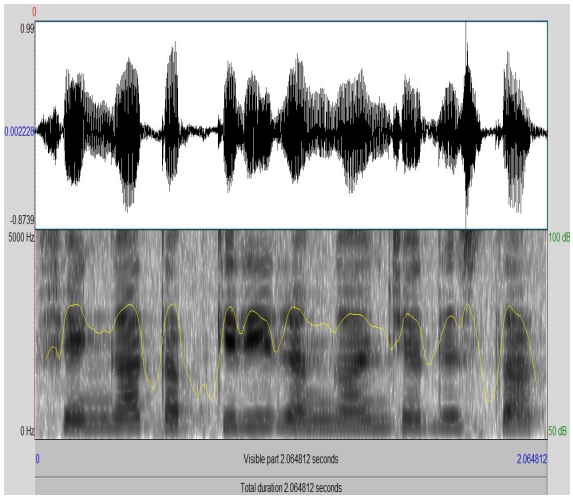

Fig. 2- Speech Waveform with Energy feature.

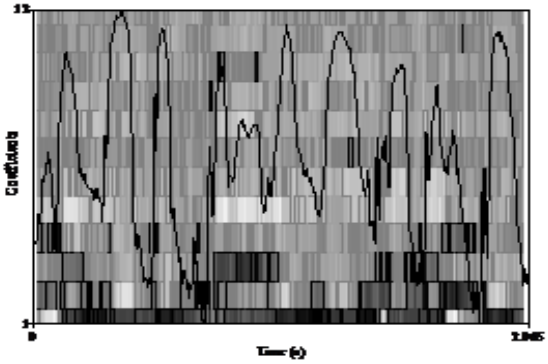

Fig. 3- Basic 12 MFCC feature.

The speech has energy as a basic dependant parameter. The energy values vary as per each frame in speech signal. The figure 2 describes that how the energy values changes as per frames in speech signal. Extraction of basic parameter from speech MFCC is robust and dynamic technique available in literature. MFCC feature values changes as per time so, figure 3 describe that the flow of MFCC parameter with respective time scale.

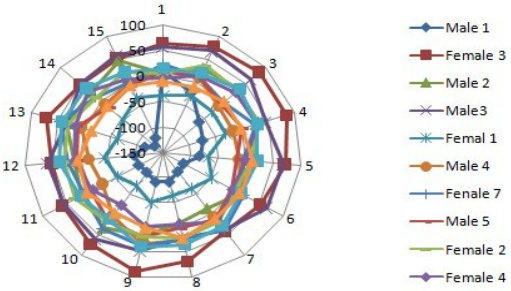

Fig. 4- Basic MFCC parameter.

We extracted basic MFCC parameter with 1first level means 12 feature. For training we selected the ten speakers sample of speech in which 05 were male and 05 were female the dependency of each male and female feature describe in figure 4. Pitch is very important independent parameter of speech, the values of pitch changes as per frame of speech that explain the figure 5.

Table 1- Recognition of the system with respective codebook size

Code Book size	Male	Female	Time (sec)
05	91.5	85.3	03
10	93.1	86.11	05
15	93.3	87.12	07
20	93.7	87.89	07
25	94.5	88.12	08

Table 2- Result of system with basic pitch parameter and threshold

Parameter	Male	Female
Mean	164.5144	202.3134
Standard deviation	23.6838	17.0531
Threshold	185.41	185.41

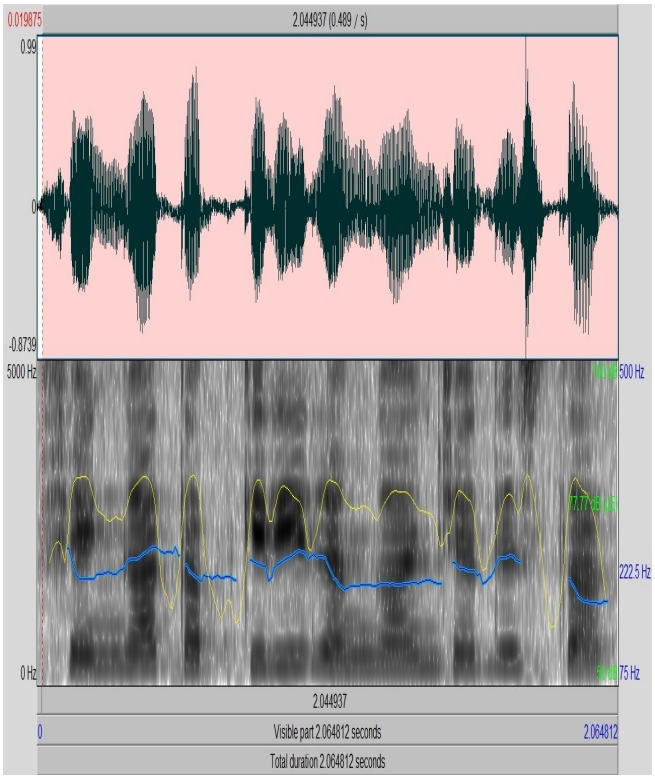

Fig. 5- Speech waveform with Pitch parameter.

The basic MFCC 12 feature with respective speaker were extracted. As well as selected values of energy parameter with respective time slice.

The Support vector machine used for clustering approach The number of test set passed to support vector machine for testing is called as codebook. The performance of the test gender identification system is on the basis of comparative codebook size the table 1 describes the recognition accuracy with respective codebook size.

Table 3- Performance of system with gender wise

Type	Accuracy (%)
Male	93.22
Female	86.90

The threshold values were needed to differentiate male and female voice. Average mean and standard deviation used to decide threshold.

The values of threshold, mean, standard deviation explain in table The performance of gender identification system calculated for main key factor

- Gender
- Age

The result changes as per the age group and is presented in table 4

Table 4- Performance of system with age group wise

Age	Accuracy (%)
20-23	89
23-25	93
25-30	95
30-40	83

Conclusion

This paper presents a voice-based gender identification system using Support vector machine in combination of MFCC. Using MFCC the basic feature extracted is combined with energy and pitch values with respective time slice for selecting the feature vector. Using SVM the feature vector classifies as per codebook size .If the codebook size is increased the accuracy of recognition also increases .We also test the performance of system on the basis of gender wise and age wise, the maximum accuracy is obtained in 25-30 age group that is 95%.

Future Work

In future we will try that our system is robust gender identification to background noise, microphone variations, and language spoken by the speaker.

Acknowledgments

The Author is thankful for university Authority for providing infrastructure and this is work is sponsored by DST under the fast track scheme.

References

[1] Tzanetakis G., Cook P.V.S. (2001) *IEEE Transactions on Speech and Audio Processing*, 10(5).
[2] Simon Haskin (1994) *Neural Networks Comprehensive Foundation*.
[3] Parris E.S., Carey M.J. (1996) *IEEE-ICASSP*, 685-688.
[4] Soma S., Sridharan S. (1997) *IEEE TENCON Speech and Image Technologies for Computing and Telecommunications,* 145-148.
[5] Hanson H. and Chuang E. (1999) *The Journal of the Acoustical Society of America*, 106, 1064.
[6] Condos A. (2004) *Digital speech: coding for low bit rate communication systems. John Wiley and Sons Ltd*.
[7] Acer and Huang X. (1996) *IEEE International Conference On Acoustic Speech and Signal Processing*.

[8] NetIQ C. and Rooks S. (1997) *IEEE workshop on Automatic Speech Recognition and Understanding*, 192-198.
[9] E. Parris and M. Carey(1996), Language independent gender identification, IEEE International Conference On Acoustics Speech and Signal Processing ,Vol 2.
[10] M. Golfer and V. Mikes (2005), The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels Journal of Voice, vol. 19, no. 4, pp. 544-554.
[11]W. Hess(1983), Pitch determination of speech signals: algorithms and devices. Springer.
[12]M. Ross, H. Shaffer, A. Cohen, R. Freud berg, and H. Manley (1974) Average magnitude difference function pitch extractor, IEEE transactions on acoustics, speech and signal processing, vol. 22, no. 5, pp. 353-362.
[13]E. Yucesoy and V. Nabiyev(2009) Gender identification of the speaker using DTW method Proceedings of the 2009 IEEE 17th Signal Processing and Communications Applications Conference, pp. 273-276.
[14]Parris E. and Carey M. (1996) *IEEE International Conference On Acoustic Speech and signal Processing,* 2.
[15]Reflection M. and Coefficients L., *Automatic Gender Identification under adverse condition*.
[16]Ting H., Yingchun Y. and W. Zhaohui (2006) 8*th International Conference*, 1.
[17]Elghonemy M., Fikri M. (2008) *IEEE International Conference on ICSSP*.
[18]Hiromi Sakaguchi, Naoaki Kawaguchi (1995) *Journal of the faculty of Engineering*, 75.
[19]Steven W. Smith (1997) *The Scientist and Engineer's Guide to Digital Signal Processing*, 169-174.
[20]Jelinek F., Bahl L.R. and Mercer R.L. (2010) *Design of a linguistic statistical decoder for the recognition of continuous*.
[21]Yanand Y., Bernard E. (1995) *ICASSP*, 3511.