Voice Gender Detection Using Gaussian Mixture Model

Prashant Kumar [1], Prashant Baheti [2], Rushab Kumar Jha [3], Preetam Sarmah [4], K.Sathish [5]
Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India
Asst. Professor, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

Abstract – In this paper, a system, developed for speech encoding, analysis, synthesis and gender identification is presented. A typical gender recognition system can be divided into front-end system and back-end system. The task of the front-end system is to extract the gender related information from a speech signal and represents it by a set of vectors called feature. Features like energy, ZCR, MFCC and Entropy. The feature is extracted using Gaussian Mixture Model (GMM) algorithm. The task of the back-end system (also called classifier) is to create a gender model to recognize the gender from his/her speech signal in recognition phase. This paper presents the processing of a speech signals which are taken from 50 people, 25 of them are Male and the rest of them are Female. The Energy and other above mentioned features of the signal is examined. The frequency at maximum power of the audio is extracted from the estimated power spectrum. The system uses the GMM technique as identification tool. The recognition accuracy of this system is 86% on average.

Index Terms – GMM,Gender,Voice,ZCR,MFCC.

1. INTRODUCTION

The human voice is comprised of sound made by a human being using the vocal chord for talking, singing, laughing, crying and shouting. It is particularly a piece of human sound creation in which the human vocal chords is the essential sound source, which plays an important role in the conversation. The application of speech or voice processing technology plays a important role in human-computer interaction. The system improves gender identification. The term gender identification deals with finding out the gender of a person from his or her voice. Gender identification has been implemented in several Automatic Speaker Recognition (ASR) systems and has proved to be of great significance. The use of gender identification in today's technology makes it easier for user authentication and identification in high security systems. There is a lot of information that can be extracted from a speech sample, for example, who is the speaker , what is the gender of the speaker, what is the language being spoken, with what emotions the speaker spoken the sentence, the number of speaker in the conversation, etc. In the field of speech analytics with machine learning, gender detection is perhaps the most foundational task. This project is dedicated toward making foray into the field of speech processing with a python implementation of gender detection from speech. We will give a brief primer about how to work with speech signals. From the speech signals in training data, a popular speech feature, Mel Frequency Cepstrum Coefficients (MFCCs), will be extracted; they are known to contain gender information (among other things).The two gender model will be by using yet another famous ML technique- Gaussian Mixture Model (GMMs).It will take input from the training samples and try to learn their distribution, which will be representative of the gender. Now, when the gender of the new voice sample is to be detected, first the MFFCs of the sample will be extracted and then the trained GMM models will be used to calculate the scores of the features for both the models .Model with the maximum score is predicted as the gender of the test speech. The common applications of these gender-dependent systems are speaker recognition and identification, multimedia annotation, and speaker indexing, annotation in multimedia, speaker recognition. Speaker diarization and speech synthesis are greatly enhanced with the application of gender identification. Other common use of gender-dependent systems include gender-wise sorting of telephone calls for surveys related to gender-sensitization, and detecting telephone call speeches from unsatisfied or angry callers.

2. OBSERVATIONS

- Having extracted the speech frames, we now proceed to derive MFCC features for each speech frame.
- Speech is produced by humans by filtering applied by our vocal tract on the air expelled by our lungs.
- The properties of the source (lungs) are common for all speakers; the properties of the vocal tract, are responsible for giving shape to the spectrum of signal and it varies across speakers.
- The shape of the vocal tract governs what sound is produced and the MFCCs best represent this shape.
- MFCCs are mel-frequency cepstral coefficients which are some transformed values of signal in cepstral domain.

From theory of speech production, speech is assumed to be convolution of source (air expelled from lungs) and filter (our vocal tract). The purpose here is to characterise the filter and remove the source part. In order to achieve our goal.

3. MOTIVATIONS

- In a mobile healthcare system, automatic gender detection can play a significant role.
- There are some vocal folds pathologies , which are biased to a particular gender; for example, vocal folds cyst can be seen particularly in female patients .
- If there is a mechanism to automatically detect the gender of the patient, it is easier for a care giver or a healthcare professional to prescribe the appropriate treatment
- The decision along with medical data can be transmitted to registered healthcare professionals for proper treatment.
- In most of the studies, the acoustic features used for the gender detection depend on the accurate estimation of the fundamental frequency.
- The accurate estimation of the fundamental frequency is itself a challenging task.

4. RELATED WORK

Gender recognition is a task of recognizing the gender from his or her voice. With the current concern of security worldwide speaker identification has received great deal of attention among of the speech researchers. Also a rapidly developing environment of computerization, one of the most important issues in the developing world is speaker identification. Speech processing based several types of research work have been continuing from a few decade ago as a field of digital signal processing (DSP).

The most efficient related work is “Speaker recognition in a multi-speaker environment” was submitted in Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001) (Aalborg, Denmark, 2001), pp. 787–90[3]. Another related work is “Spectral Feature for Automatic Voice-independent Speaker Recognition” was developed in the department of Computer Science, Joensuu University, Finland 2003. From the study of different previous research works it was observed that among the different features GMM and FFT provide the results in best classification rate.

Based on the GMM Model, we have computed frequency spectrum from maximum power of speech signal. We have implemented a complete gender recognition system to identify particular gender (male/female) using frequency component. In addition to description, theoretical and experimental analysis, we provide implementation details as well.

5. WORKING

A speech signal is just a sequence of numbers which denote the amplitude of the speech spoken by the speaker. We need to understand 3 core concepts while working with speech signals:

5.1 FRAMING

Since speech is a non-stationary signal, its frequency contents are continuously changing with time. In order to do any sort of analysis of the signal, such as knowing its frequency contents for short time intervals (known as Short Term Fourier Transform of the signal), we need to be able to view it as a stationary signal. To achieve this stationaryty , the speech signal is divided into short frames of duration 20 to 30 milliseconds, as the shape of our vocal tract can be assumed to be unvarying for such small intervals of time. Frames shorter than this duration won’t have enough samples to give a good estimate of the frequency components, while in longer frames the signal may change too much within the frame that the condition of stationary no more holds.

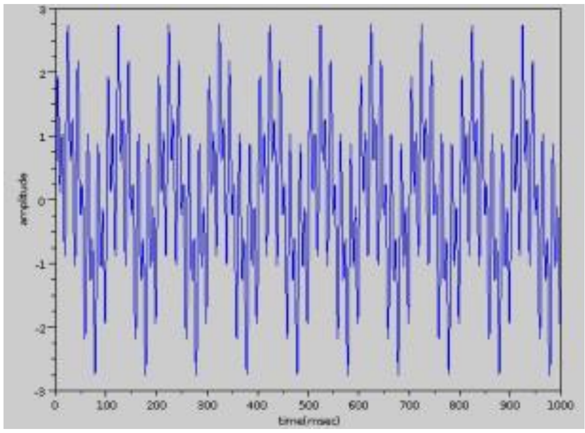

Figure 1 Full Signal

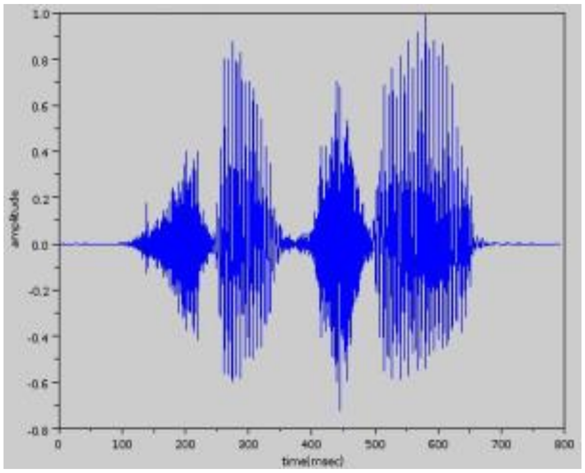

Figure 2 Processed Signal

5.2 WINDOWING

Extracting raw frames from a speech signal can lead to discontinuities towards the endpoints due to non-integer number of periods in the extracted waveform, which will then lead to an erroneous frequency representation (known as spectral leakage in signal processing lingo). This is prevented by multiplying a window function with the speech frame. A window function’s amplitude gradually falls to zero towards its two ends and thus this multiplication minimizes the amplitude of the above mentioned discontinuities

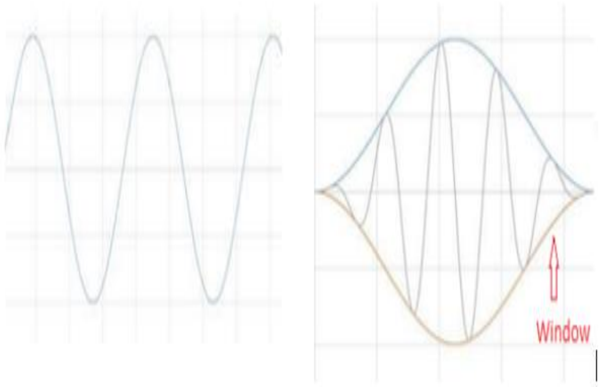

Figure 3 Windowing

5.3 OVERLAPPING FRAMES

Due to windowing, we are actually losing the samples towards the beginning and the end of the frame; this too will lead to an incorrect frequency representation. To compensate for this loss, we take overlapping frames rather than disjoint frames, so that the samples lost from the end of the i th frame and the beginning of the (i+1)th frame are wholly included in the frame formed by the overlap between these 2 frames. The overlap between frames is generally taken to be of 10-15 ms

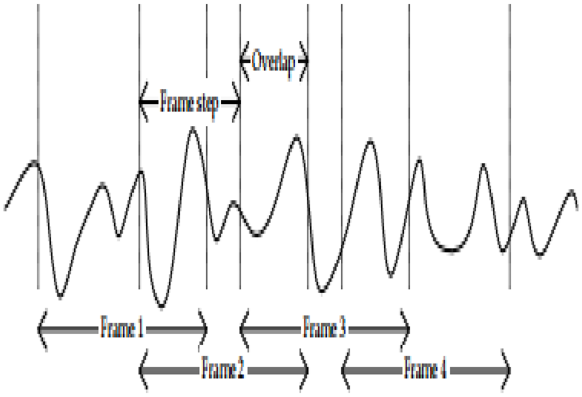

Figure 4 Overlapping Frames in a Speech Signal

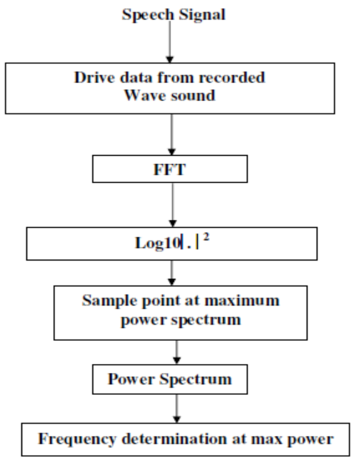
Speech Signal
Drive data from recorded Wave sound
FFT
Log10| . |^2
Sample point at maximum power spectrum
Power Spectrum
Frequency determination at max power

Figure 5 Block Diagram

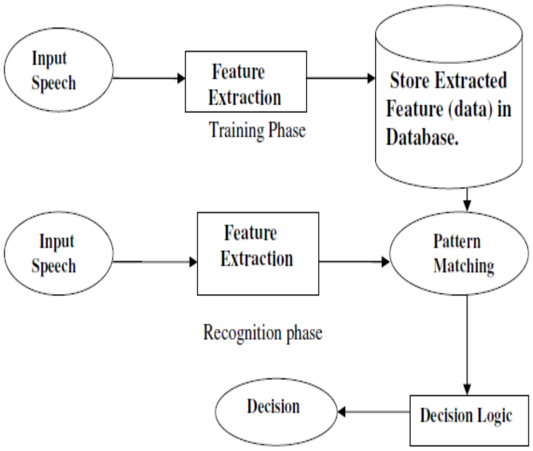

Figure 6 System Architecture

5.4 EXTRACTING MFCC FEATURE

Having extracted the speech frames, we now proceed to derive MFCC features for each speech frame. Speech is produced by humans by filtering applied by our vocal tract on the air expelled by our lungs. The properties of the source (lungs) are common for all speakers; it is the properties of the vocal tract, which is responsible for giving shape to the spectrum of signal and it varies across speakers. The shape of the vocal tract governs what sound is produced and the MFCCs best represent this shape.

MFCCs are mel-frequency cepstral coefficients which are some transformed values of signal in cepstral domain. From theory of speech production, speech is assumed to be convolution of source (air expelled from lungs) and filter (our vocal tract). The purpose here is to characterise the filter and remove the source part. In order to achieve this,

- We first transform the time domain speech signal into spectral domain signal using Fourier transform where source and filter part are now in multiplication.
- Take log of the transformed values so that source and filter are now additive in log spectral domain. Use of log to transform from multiplication to summation made it easy to separate source and filter using a linear filter.
- Finally, we apply discrete cosine transform (found to be more successful than FFT or I-FFT) of the log spectral signal to get MFCCs. Initially the idea was to transform the log spectral signal to time domain using Inverse-FFT but 'log' being a non-linear operation created new frequencies called Quefrency or say it transformed the log spectral signal into a new domain called cepstral domain (ceps being reverse of spec).
- The reason for the term 'mel' in MFFC is mel scale which exactly specifies how to space our frequency regions. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear.

5.5 TRAINING GENDER MODULE

In order to build a gender detection system from the above extracted features, we need to model both the genders. We employ GMMs for this task.

A Gaussian mixture model is a probabilistic clustering model for representing the presence of sub-populations within an overall population. The idea of training a GMM is to approximate the probability distribution of a class by a linear combination of 'k' Gaussian distributions/clusters, also called the components of the GMM. The likelihood of data points (feature vectors) for a model is given by following equation:

$$P(X|\lambda) = \sum_{k=1}^{K} w_k P_k(X|\mu_k, \Sigma_k) \quad (1)$$

, where $P_k(X|\mu_k, \Sigma_k)$ is the Gaussian distribution

$$P_k(X|\mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi|\Sigma_k|}} e^{\frac{1}{2}(X-\mu_k)^T \Sigma^{-1}(X-\mu_k)} \quad (2)$$

The training data X_i of the class λ are used to estimate the parameters mean μ, co-variance matrices Σ and weights w of these k components.

Initially, it identifies k clusters in the data by the K-means algorithm and assigns equal weight $w = \frac{1}{k}$ to each cluster. k Gaussian distributions are then fitted to these k clusters. The parameter'sμ, σ and w of all the clusters are updated in iterations until the converge. The most popularly used method for this estimation is the Expectation Maximization (EM) algorithm.

Python's sklearn.mixture package is used by us to learn a GMM from the features matrix containing the MFCC features. The GMM object requires the number of components n_components to be fitted on the data, the number of iterations n_iter to be performed for estimating the parameters of these n components, the type of co-variance covariance_type to be assumed between the features and the number of times n_ init the K-means initialization is to be done. The initialization which gave the best results is kept. The fit() function then estimates the model parameters using the EM algorithm.

5.6 EVALUATION OF SUBSET OF AUDIOSET

Upon arrival of a test voice sample for gender detection, we begin by extracting the MFCC features for it, with 25 ms frame size and 10 ms overlap between frames . Next we require the log likelihood scores for each frame of the sample, $x_1, x_2, \ldots, x_i$, belonging to each gender, ie, $P(x_i|female)$ and $P(x_i|male)$ is to be calculated. Using (2), the likelihood of the frame being from a female voice is calculated by substituting the μ and Σ of female GMM model. This is done for each of the k Gaussian components in the model, and the weighted sum of the k likelihoods from the components is taken as per the w parameter of the model, just like in (1). The logarithm operation when applied on the obtained sum gives us the log likelihood value for the frame. This is repeated for all the frames of the sample and the likelihoods of all the frames are added.

Similar to this, the likelihood of the speech being male is calculated by substituting the values of the parameters of the trained male GMM model and repeating the above procedure for all the frames. The Python code given below predicts the gender of the test audio.

6. RESULT

The program shows the results of the evaluation on the subset extracted from AudioSet corpus. The approach performs brilliantly for the female gender, with an accuracy of 95%, while for the male gender the accuracy is 76%. The overall accuracy of the system is 86%.

If we look at the performance, trained female model seems to be good representative of their gender in comparison to trained male model. We can also evaluate the same trained gender models on an exhibition data-set that consisting of more data set increasing the accuracy of the model. Therefore meaning that the accuracy of the program increases with the increase in the dataset,that is used to train the program.

7. CONCLUSION

By using this model we can easily predict the gender of an individual person and use it in an effective way to work such as in healthcare domain, easy banking, tracking or keeping audio records, security purposes and in many other ways that can save time as well as increase the efficiency of the work.

REFERENCES

[1] Prabhakar, S., Pankanti, S., and Jain, A. "Biometric recognition: security and privacy concerns" IEEE Security and Privacy Magazine 1(2003), 33-42.

[2] Huang X., Acero, A., and Hon, H.-W. "Spoken Language Processing: a Guideto Theory, Algorithm and System Development" prentice-Hall, New Jersey, 2001.

[3] Martin, A., and Przybocki, M. Speaker recognition in a multi-speaker environment.In Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001) (Aalborg,Denmark, 2001), pp. 787–90.

[4] Tomi Kinnunen " Spectral Feature for Automatic Voice-independent Speaker Recognition" Depertment of Computer Science, Joensuu University,Finland. December 21, 2003.

[5] John R. Deller, John G Proakis and John H. L. Hansen, "Discrete- Time Processing of Speech Signals" Macmillan Publishing company, 866 Third avenue, New York 10022.

[6] Rabiner Lawrence, Juang Bing-Hwang, "Fundamentals of Speech Recognitions", Prentice Hall New Jersey, 1993, ISBN 0-13-015157-2.

[7] Md. Saidur Rahman, "Small Vocabulary Speech Recognition in Bangla Language", M.Sc. Thesis, Dept. of Computer Science & Engineering, Islamic University, Kushtia-7003, July-2004.

Authors

Prashant Kumar
Undergraduate Student
Department of Computer Science,
SRM Institute of Science and Technology, Chennai-600089

Prashant Baheti
Undergraduate Student
Department of Computer Science,
SRM Institute of Science and Technology, Chennai-600089

Rushab Kumar Jha
Undergraduate Student
Department of Computer Science,
SRM Institute of Science and Technology, Chennai-600089

Preetam Sarmah
Undergraduate Student
Department of Computer Science,
SRM Institute of Science and Technology, Chennai-600089

K.Sathish
Assistant Professor (O.G)
Department of Computer Science,
SRM Institute of Science and Technology, Chennai-600089