

LOW LEVEL SPEECH FEATURES BASED GENDER CLASSIFICATION OF A SPEAKER

¹RAJNIKANT P SANDHANI, ²RAVI G MEHTA

^{1,2} EC Dept. , Govt. Engineering College, Rajkot

sandhanirp@gmail.com

ABSTRACT:

Gender classification present significant challenges in speech processing. A crucial aspect involves feature selection from the speech data. Pitch is commonly utilized in speech processing for distinguishing between genders, as male and female speech exhibit distinct pitch ranges. Also many other features are available from speech of a person which can be from temporal class or from spectrum. Only time domain or frequency domain features based researches are mostly considered. Here effort have been made to do gender classification based on both the category. Many clustering and classification methods also available for this kind of task. SVM only is considered for classification as a limited case. Samples of persons were taken from less noisy and heavy noise environment. Over all good performance seen from spectral case than from the temporal. Our experiment have concluded to have more accuracy in female than male speakers.

KEY WORDS: Pitch, Formants, zero crossings mean (ZCM), average amplitude, RMS energy, Mel frequency cepstral coefficients (MFCC), Linear frequency cepstral coefficients (LFCC)

1. INTRODUCTION

Gender classification based on speech typically involves analyzing various acoustic and linguistic features to predict the gender of a speaker. Pitch can be measured in terms of fundamental frequency (F0). Men generally have lower average pitch frequencies compared to women. Formants are resonant frequencies in the vocal tract that differ between genders due to physiological differences [1][2][4][5]. Formant frequencies are typically higher in women than in men. Intensity or loudness is slightly higher in men voice than women on average. Women often speak at a faster rate than men. Men and women may exhibit different patterns of pauses during speech. Gender classification systems typically use machine learning algorithms to analyze speech features and make predictions. Supervised learners like Support Vector Machines (SVM), Decision Trees, or Neural Networks etc. can be trained on labelled datasets where the gender of speakers is known [8][9].

Speaker Variability is the Individual differences in speech can sometimes be larger than gender differences, making accurate classification challenging. Speech patterns can vary widely across cultures and social groups, which may affect the accuracy of gender classification. Choosing which features and how to represent them (e.g., frequency domain, time domain) is crucial for the effectiveness of the system. Gender recognition based on speech can have various applications, such as in forensic analysis (e.g., identifying speakers from audio recordings), voice-controlled systems (e.g., personal

assistants), and market research (e.g., analyzing consumer preferences based on gender).

Overall, gender classification based on speech is a multidimensional task that involves sophisticated analysis of features and utilizing advanced machine learning techniques to achieve accurate predictions.

2. SYSTEM OVERVIEW

Preprocessing of the speech is important part of the front end speaker recognition system which involve noise immunity, uniformity in spectrum and much more. In 5-10msec, speech characteristics does not varies significantly hence short time spectral analysis is possible. We know that, characteristic voice of a person is affected by person to person due to variation in vocal tract length, vocal cords size, nasal cavity size and other related organs of the articulatory system [3][4][5]. Hence formant frequency, average pitch, pitch range etc. in the spectral information may be affected. Age and health of the person also plays important role. Pitch frequency decreased with increase in the age was reported by some experiments. Health condition like cold may produce sound more from nasal.

For collecting parameters of the speech, frame size of around 25ms can be taken and it is in step of 10ms with 15ms of overlapping between frames are chosen to extract features. Low level feature like MFCC which is directly based on frame are extracted. Preprocessing is to enhance the high frequency part of spectrum for uniformity in loudness and then cepstral features are obtained [4]. Speech is passed through window and then squared to be useful for Mel filter bank. Mel filter bank is chosen be matched

with human auditory scale and shaped in triangular. Near to 25 filters are used in banks for this purpose then DCT gives the cepstral coefficients [5]. Many frequency wrapping methods available for the generation of cepstral features [4]. Mel frequency cepstral coefficients (MFCC) are mostly used in speech and speaker recognition task due somewhat frequency scale matching to human ear process on sound [6][7]. Here MFCC filter banks used have bandwidth increasing progressively from start to end and hence have good frequency resolution in lower part of spectrum and have less frequency resolution in upper end of spectrum. As per the voice production mechanism theory according to length of vocal tract some formants position change particularly on higher side. Vocal tract length have difference around 2-3 cms for male and female case [3]. Hence there may be more variation between cases of different gender. As MFCC is having good frequency resolutions in lower frequency bands than higher bands, some useful information in higher frequency spectrums may be omitted due the this limitation of MFCC[10]. So MFCC works well for male compared to female recognition scenario. Linear frequency cepstral coefficients (LFCC) have all equal bands of frequencies and uniformly distributed along spectrum region of interest. But in many researches it was reported that MFCC was sometime better performs than LFCC. In [10], for female and babble noise, LFCC was found to be better performed. So, there should be selection of frequency wrapping method based on the gender type to give good result of speaker recognition. Fig. 1 shows the two phases in gender recognition system, one for training the system and then second for actual gender recognition[8][9].

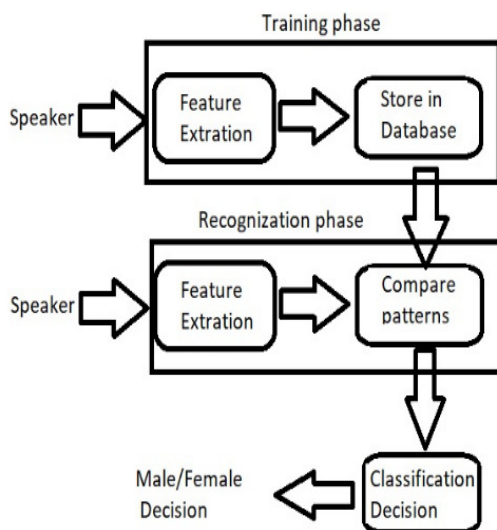


Fig.1 Basic functional block of gender classification system

3. FEATURES DESCRIPTION

In most cases of speech processing, speech features extraction plays important role. Here some features are briefly explained in this section. The pitch period shows the duration between consecutive voiced excitation cycles, specifically from one peak to the next. It is the fundamental frequency of the speech excitation source.

Formant features are essentially selective, non-uniform samples of the signal spectrum positioned at the resonance frequencies of the vocal tract. These frequencies typically exhibit superior signal-to-noise ratios compared to other spectral components. It can be visible from the fig.2.

Zero crossing rate in speech processing is a measure of how often the signal changes polarity (crosses the zero axis) within short frames of audio. It provides valuable information about the temporal characteristics of the signal. Here mean of zero crossings is used.

RMS (Root Mean Square) energy in speech processing refers to a measure of the energy present in a segment of speech signal. It can be computed by squaring the mean value of the squared values of the frame samples in a segment of the speech signal.

Average amplitude in speech processing refers to the mean magnitude of the speech signal within a segment or time window, providing valuable information about the signal's intensity and aiding in various speech analysis tasks. It can be visible from the fig.3.

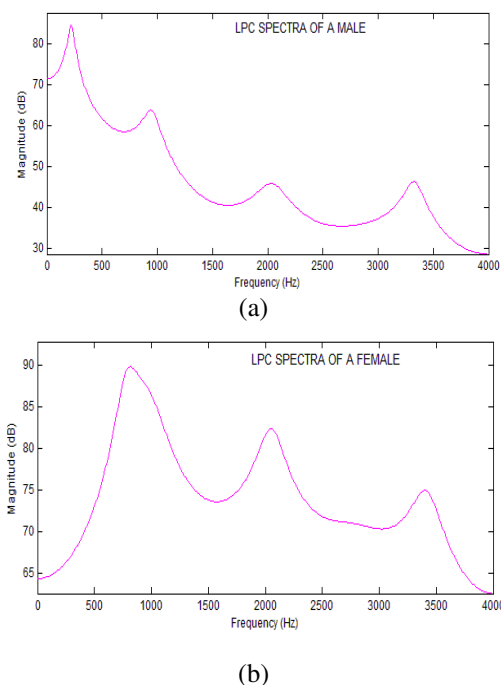


Fig.2 LPC spectra of a (a) male and (b) female (By Colea)

MFCCs are a powerful tool in speech processing that transforms the complex speech waveform into a compact representation of the spectral characteristics of speech, modeled after human auditory perception. Their effectiveness and robustness have made them a standard feature extraction technique in many speech-related applications.

LFCC stands for Linear-frequency cepstral coefficients. Similar to MFCCs, LFCCs are another type of feature extraction technique used in speech processing. MFCCs utilize a Mel-scale filter bank to approximate the human auditory system's frequency response, while LFCCs use a linear-scale filter bank.

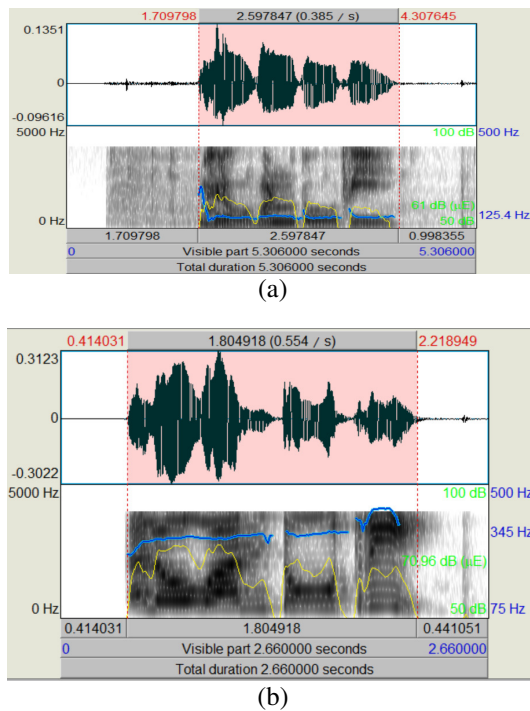


Fig.3 Spectrogram, intensity and pitch of a (a) male and (b) female (By PRAAT)

4. EXPERIMENTAL DATA USED

For our experimentation we have taken total 60 samples of persons in less noisy environment. From that 35 belong to male and 25 belongs to female. In this, 3 females have voice as looks like male and 1 male having voice looks like female. For the case of noisy environment, separate voice samples of 6 male and 4 female were collected in city area road where vehicle noise was also significant. Speakers of all samples were free to speak anything not specific words. All the samples were truncated to 2 seconds and various speech features were extracted from that based on the average values over all speech segments for that speaker. COLEA and PRAAT tools were utilized for this experimentation. For time domain features we have used the zero crossings mean (ZCM), average amplitude and RMS energy. In frequency domain case we have used the pitch, first

four formants, LFCC and MFCC. Distance measured only by the Euclidean distances. For the classification purpose SVM considered. Samples allocated for training is 60% and that are for testing is 20%. Remaining 20% used for measuring the actual accuracy of the system. Noisy samples are also used to test the accuracy of system.

Accuracy is measured for true recognition of gender from the total count available for that gender under experiment. Possible major errors in recognition are of two type. One can accept imposter speaker as correct one which is known as false alarm and in second it reject true speaker called false reject or missed detection. Errors may occur due to some extrinsic factors like noise, condition of channel and some intrinsic factors like style of speaking, emotion, gender, age etc..

5. PERFORMANCE OF EXPERIMENT

It is found to have overall good accuracy by ZCM among time base features. RMS energy and average amplitude have significant less accuracy may be due to loudness variation from speaker to speaker as in The pitch period shows the duration between consecutive voiced excitation cycles, specifically from one peak to the next. It is the fundamental frequency of the speech excitation source as shown in Fig.3. Female recognition accuracy found more compared to male cases in this experimentation based on time features. Table 1 gives accuracy for time based features.

From frequency based features recognition, it is seen that pitch and format based classification have more accuracy in case of female compared to male. In comparison of accuracy from MFCC and LFCC, it is found that female have more accuracy in LFCC. This is due to more high frequency features concentration in female speech were detected precisely by linear scale compared to Mel scale where less frequency resolution is observed in higher frequency region. Table 2 gives accuracy of frequency based features. When speaker having cross gender wise similar voice, much less accuracy and more confusion generated many times. For the noisy samples test it shows overall 5-10% less performance for both the genders. Frequency based features found more robust compared to time domain features.

Table 1: Accuracy of time domain features in less noisy environment

Accuracy--> ↓ Features	% Accuracy for Male	% Accuracy for Female
ZCM	57	60
Avg. Amp.	49	52
RMS energy	51	56

Table 2: Accuracy of frequency domain features in less noisy environment

Accuracy--→ ↓ Features	% Accuracy for Male	% Accuracy for Female
Pitch	63	64
Formants	66	68
LFCC	60	68
MFCC	69	64

cepstral coefficients for speaker recognition” In proceeding Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, 2011, 559-564.

6. CONCLUSION AND FUTURE SCOPE

In this paper gender classification based on temporal and spectral features is evaluated. Female classification found more accuracy than male in overall scenario may be due to tonal difference between them. Zero crossing have good accuracy compared to other features used due to normalization issue in other features. Further experiments can be done considering all aspects of front end preprocessing and more robust clustering and classification algorithm.

7. REFERENCES

- [1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. ASSP*, vol. 28, no. 4, 357-366, 1980
- [2] K. N. Stevens, *Acoustic phonetics*. Cambridge, Mass.: MIT Press, 1998.
- [3] Story, B.H., (2003). "Using imaging and modeling techniques to understand the relation between vocal tract shape and acoustic characteristics", *Proceedings of the Stockholm Music Acoustics Conference*, 6-9 August.
- [4] L.R. Rabiner, Juang, Yegnanarayana, "Fundamentals of Speech Recognition", Pearson Education, New Delhi, India, 2009.
- [5] L.R. Rabiner and R. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Englewood Cliffs, New Jersey, 1979.
- [6] Lara Lynn Stoll, "Finding Difficult Speakers in Automatic Speaker Recognition", PHD Thesis, Engineering - Electrical Engineering and Computer Sciences, University of California at Berkeley, December 16, 2011.
- [7] Joseph P. Campbell, "A Tutorial on Speaker Recognition", *Proceeding of the IEEE*, Vol. 85, No. 9, 1437-1462, Sept. 1997.
- [8] Zahi N. Karam, William M. Campbell "Variability Compensated Support Vector Machines applied to speaker verification", MIT Lincoln Laboratory, Lexington MA MIT, Cambridge MA, USA, September, 2009.
- [9] Kofi A. Boakye, "Speaker Recognition in the Text-Independent Domain Using Keyword Hidden Markov Models", University of California at Berkeley, May 11, 2005.
- [10] Zhou, X.; Garcia-Romero, D.; Duraiswami, R.; Espy-Wilson, C. & Shamma, S. "Linear versus mel frequency