GENDER-DEPENDENT FEATURE EXTRACTION FOR SPEAKER RECOGNITION

Lantian Li and Thomas Fang Zheng∗

Center for Speech and Language Technologies, Division of Technical Innovation and Development
Tsinghua National Laboratory for Information Science and Technology
Center for Speech and Language Technologies, Research Institute of Information Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

∗ Corresponding Author: fzheng@tsinghua.edu.cn

ABSTRACT

Gender information is believed helpful for speaker recognition. In GMM-UBM based speaker recognition, the use of gender information is mostly on a basis of constructing gender-dependent (GD) UBM s instead of extracting GD features. However, theoretical analysis and experimental observations show that females and males differ quite a lot in the feature domain of speech signal, including F0, formant, spectrum and cepstrum. In this paper, further analysis and experiments have been done to explore the differences between females and males. Afterwards, a GD MFCC feature extraction is proposed. In this method, the frame length of MFCC extraction is gender dependent, in other words the resolution for both the DFT analysis and hence the MFCC feature extraction is gender dependent. Experimental results demonstrate that compared with the gender-independent (GI) feature extraction, the GD feature extraction can achieve relative EER reductions of 21.7% and 12.2% for female and male speakers evaluations, respectively.

Index Terms— Speaker recognition, gender-dependent feature extraction, GMM-UBM, MFCC

1. INTRODUCTION

Over the last decade, the Gaussian mixture model-Universal background model (GMM-UBM) [1] and i-vector [2] have been popular approaches in speaker recognition. Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coefficients (LPCC) have been two most commonly used features in both GMM-UBM and i-vector based speaker recognition. In practice, a unified set of feature extraction configurations are defined beforehand, just like a speaker-independent feature extraction is applied. However, the acoustic characteristics differ from person to person, and this clearly violates the principle of feature extraction that is to convey speaker-dependent information. However, extracting feature with speaker-dependent configurations is unrealistic because it is almost impossible to define the configuration for every single speaker.

From the gender’s point of view, there are many differences between females and males either in speech production or acoustical characteristics, such as the structure of vocal tract, fundamental frequency (F0), and pitch [3]. Meanwhile, for a certain gender, there exists certain similarities and cohesiveness in acoustical characteristics. Therefore, the study of exploring the discrimination of acoustical characteristics in view of gender and extracting gender-dependent feature is of great significance.

Back in 1991, [4] proposed fine analysis of the characteristics of vowels, including formant frequencies, bandwidths, amplitudes, as well as speaker's fundamental frequency (F0) or pitch of voicing for gender recognition from speech. The results suggested that redundant gender information was imbedded in F0 and vocal tract resonance characteristics. In addition, researchers have explored that F0 values for males are lower than those for females due to longer and thicker vocal folds. F0 for males is typically around 110 Hz in the range of 100-146 Hz, while females around 200 Hz in 188-221 Hz [5, 6]. Furthermore, males exhibit lower formant frequencies than females due to vocal tract length differences. These differences are embedded in the LPCC extraction, which can well model the vocal tract properties by using an all-pole model and emphasize the formant structure [7, 8]. Therefore, in the light of characteristic differences between females and males, the vocal track can be modeled for males and females separately in order to estimate more accurate gender dependent features.

Based on these discoveries, researchers have attempted to apply them in speech/speaker recognition. In [9], they first performed gender classification based on the formant information in order to turn a gender-independent (GI) problem into gender-dependent (GD), and then to use the gender dependent method to perform speech recognition. [1] proposed to train two GMMs, one for males and another for females, and then to pool the two GMMs together to create a UBM for speaker recognition. To some extent, this approach solved the UBM biases when speech was unbalanced between males and females.
Lots of model-level approaches have been proposed to address the GI/GD issues and the experimental results showed that some improved GI systems using the discriminative training have achieved almost the same performance as similar GD systems [10, 11], but in the feature-level few research has been reported. In this paper, following the theoretical knowledge mentioned above, we will analyze the extraction procedure of the filterbank-based cepstral features MFCC and explore the regularity on gender dependent feature extraction.

MFCC, which can closely approximate the human auditory system's response, has been the dominant feature used in both speech recognition and speaker recognition. MFCC represents the spectral components of the signal by performing discrete Fourier transform (DFT) and further carrying out filtering based on the perceptually motivated Mel scale. Assuming that the speech signal has the short-term stationarity, windowing and en-framing are applied to the beginning of the signal and then moved further till the end of the signal is reached. Two quantities have to be determined: the frame size (or frame length) $FL$, and the frame shift (or step size) - the shift between two consecutive frames.

As shown in Eq. (1), once the sample rate $f_s$ (in Hertz) and the frame length $FL$ (in second) has been determined, the window duration / size $N$ (in point) of DFT is also determined:

$$N = FL \cdot f_s$$  \hspace{1cm} (1)

For computational purpose, DFT is performed so that the frequency variable only takes on $N$ discrete values, as shown in Eq. (2).

$$X_s(k) = \sum_{n=0}^{N-1} x(m) e^{-j2\pi km/N} w(n-m)$$  \hspace{1cm} (2)

where $x(m)$ is a speech signal, $X_s(k)$ is the DFT of $x(m)$ on time $n$, and $w(m)$ is a $N$-sample window function. The Fourier transform of $X(m)\cdot w(n-m)$ is the windowed version of $x(m)$ using a window shifted to a time $n$ with respect to the speech.

Obviously, the number $N$ determines the analysis resolution of the DFT calculation, and the choice of $N$ is crucial for short-time Fourier analysis. Besides, $N$ and frequency resolution $\Delta f$ have an inverse relationship, as shown in Eq. (3).

$$\Delta f = f_s / N$$  \hspace{1cm} (3)

In order to achieve a higher frequency resolution, the window size $N$ should be large enough, which however leads to poor time resolution and hence does not well satisfy the assumption of the short-term stationarity. On the other hand, a smaller value of $N$ corresponds to a smaller frequency resolution since the window low-pass filter is wide, but it leads to a higher time resolution since the speech properties are averaged only over short time intervals [12].

From the analysis above, it can be easily seen that the window size $N$ plays a key role in feature extraction. And if the sample rate $f_s$ is fixed, the frame length $FL$ is the only one factor that determines the value of $N$.

The step size is another important parameter, used to describe the overlapped extent between two consecutive windows in order to meet the short-time stationarity of speech signals. For a speech signal, if the step size is too big, the number of frames will be too few and there is no significant relationship between two consecutive windows. On the contrary, much-overlapped frames lead to the information redundancies among features. In current speaker recognition, the step size is often taken as about 10 ms, which is also what we take in this paper.

If the frame length $FL$ is not appropriately chosen, it will cause harmonic leakage and spectrum fuzzy, and then lead to signal information loss and superposition distortion. To overcome the defects of fixed frame-size features, pitch synchronization analysis methods were proposed in [13, 14]. The premise of pitch synchronization analysis for dynamic frame-size is to accurately capture the pitch of speech signal. These methods can be divided into three steps in general: firstly to obtain pitch information, secondly to perform dynamic framing, and thirdly to use dynamic frame-sizes to extract features. While the experimental results show that using this kind of methods can improve the system performance, but there are still many weaknesses. In fact, the pitch $P_T$ is person-specific and its value range is very wide, thus how to find the relationship between the frame length $FL$ and pitch period $P_T$ is uncertain. Meanwhile, using the pitch synchronization analysis to dynamically generate the frame length lowers down the efficiency of feature extraction process, and it is no conducive to practical application.

Overall, in the state-of-art speaker recognition system based on GMM-UBM or i-vector, researchers usually choose the frame length and step size empirically or experimentally, yet do not explore the regularity on it. In this paper, from the gender’s perspective, we explore the regularity of frame-size-impact on GD speaker recognition and expect to explain the discrimination between males and females in speech production and acoustical characteristics so as to propose a GD MFCC feature extraction method.

The remainder of this paper is organized as follows. The GD feature extraction method is given in Section 2. In Section 3, experiments are described with results and analysis presented. Conclusions are drawn in Section 4.

### 2. THE GENDER-DEPENDENT FEATURE EXTRACTION

As mentioned in Section 1, many researches have been conducted to explore the differences between females and males either in speech production or acoustical characteristics but lots of work have been done to address the GI/GD issue. However, generally in these previous
studies, no matter for GI or GD task, features are extracted with GI configurations. Obviously, it does not take gender information into consideration for feature extraction. To further study this situation, in this paper, we explore the differences between males and females in speech production and acoustical characteristics, and focus on proposing a gender-dependent feature extraction method.

The framework of the proposed GD feature extraction is shown in Fig. 1.

There are two key parts: one is the feature extraction process, which is to evaluate the performance as a function of frame length so as to extract several groups of MFCC features. Another one is to apply these different groups of features in either GD or GI speaker recognition systems based on the GMM-UBM structure.

![Fig. 1 Block diagram of GI/GD feature extraction experiments](image)

**2.1. The GMM-UBM based speaker verification**

In this paper, the popular GMM-UBM framework based speaker verification system is used as detailed in [1].

**2.2. Finding the optimal analysis resolution**

In order to find the optimal analysis resolution $FL$, the first step is to observe the effect of the frame length $FL$ on the GMM-UBM based speaker verification. Therefore, we choose different values of $FL$ varying from 12 ms to 48 ms with the interval of 4 ms in gender-independent MFCC extraction. Note that we assume the step size is only associated with the number of frames, so it is chosen as a constant value of 10 ms. In a word, experiments are designed to verify the theoretical analysis in Section 1 and we expect to observe the system performance as a function of frame length $FL$ based on the GI speaker recognition system.

In the second step, experiments are done on GD feature extraction to find the optimal $FL$ values for females and males, respectively. There are two evaluations, one using the female development set only and another using the male development set only. For each evaluation, we choose a specific frame length, accordingly perform feature extraction, generate a GD UBM, adapt speaker models, and then evaluate the system performance based on the corresponding evaluation set, finally we can obtain a performance curve as a function of frame length.

**3. EXPERIMENTS AND ANALYSIS**

**3.1. Database**

The database used is called CSLT-DSDB (Digit String Database) that was jointly created by CSLT (Center for Speech and Language Technologies), Tsinghua University and Beijing d-Ear Technologies, Co. Ltd. The recording was conducted using different mobile microphones, sampled at 16 kHz with 16-bit precision.

The database is composed of two parts:

1) Development set: used for UBM training with an approximate size of 1 GB data (about 200 males and 200 females) recorded in an ordinary office environment.

2) Evaluation set: for purpose of the speaker enrollment and verification, consisting of 224 speakers’ data (116 males and 108 females). For each speaker, there are text-prompted digit strings of about 40 seconds in length for speaker model training; and 8~12 randomly generated digit strings each of which is an 8-digit string (of about 1~2 seconds in length) for verification.

**3.2. Observations**

The first thing to do in our experiments is to verify the theoretical analysis in Section 1 and observe the effect of frame length on the GI system. First of all, the overall development set (consisting of all development set) is used to train an overall UBM, and accordingly the evaluation is performed to get an overall result for the GI case (called Overall Evaluation). Then, experiments based on the female evaluation set and the male evaluation set are performed using the same GI-UBM, called Female Evaluation and Male Evaluation, respectively. The verification performance is evaluated in terms of the popular Equal Error Rates (EER). Results are presented in Fig. 2.

![Fig. 2 Evaluations with UBM trained using all development sets](image)

From the gray dotted line based on overall evaluation set, we observe that the frame length has a great influence on the system performance and this EER curve has the same trend with a parabola going upwards. We also confirm that this curve is consistent with our analysis that an
inappropriate value of frame length $FL$ (either too big or too small) can result in system performance degradation. The best performance is achieved when the frame length $FL$ equals to 20 ms, 20 ms, and 16 ms with the EER of 0.99%, 0.90% and 1.06% for overall, male, and female evaluations, respectively.

### 3.3. Gender-Dependent Feature Extraction

Observations in Section 3.2 reveal that the frame length $FL$ is an important parameter in feature extraction. Besides, as mentioned in Section 1, there are many differences between females and males in speech production and acoustical characteristics, where $FL$ can reflect these differences to some extent. Experiments are then done to verify our hypothesis, following the descriptions in Section 2.2.

For female-specific evaluations, given a predefined frame length, we perform feature extraction, generate a UBM using the whole female development set (named Female-UBM), adapt speaker models, and then evaluate the system performance across both the female evaluation set (gender-matched case) and the male evaluation set (gender-mismatched case), by varying the frame length we can obtain a performance curve as a function of frame length as in Fig. 3. Similar evaluations are performed for males and results are plotted as in Fig. 4.

![Fig. 3 Evaluations with UBM trained using female development set](image)

![Fig. 4 Evaluations with UBM trained using male development set](image)

We find that the solid line in Fig. 3 and the dotted line in Fig. 4 show the same trend as in Fig. 2, which are the gender-matched cases between training and evaluation. We also observe that the gender-matched cases outperform the gender-mismatched cases quite a lot. It is most possibly due to the differences of feature distributions between females and males, which lead to the difference of the spatial distribution between female-UBM and male-UBM.

For either female-specific evaluations or male-specific evaluations, it can be seen that a minimal EER can be achieved at a certain frame length condition yet the minimal-EER frame lengths are different for females and males. For female-specific evaluations, the minimal EER is achieved when the frame length $FL$ equals to 20 ms while for male-specific evaluations, $FL$ equals to 24 ms or 28 ms. It is consistent with the analysis in Section 1 that compared with males, the pitch durations of females are shorter. So $FL$ should be shorter for female than male in order to achieve a higher frequency resolution and extract more accurate features.

Compared with the minimal EERS of female and male evaluation curves shown in Fig. 2, this GD feature extraction approach can achieve relative EER reductions of 21.7% and 12.2% for female and male evaluation, respectively. For GI feature extraction, almost all experiments conclude that female evaluations achieve worse performance than the male evaluations. By utilizing the GD feature extraction, performance on female data can be much more improved and best performances for female and male are almost balanced. Although the regularity of this GD feature extraction is not apparent in gender-matched evaluation cases, we believe the observations and conclusions above are still meaningful and extensible.

### 4. CONCLUSIONS

In this paper, we propose a GD feature extraction solution to speaker recognition. Experimental observations show that to get better speaker recognition performance, the optimal frame length for females should be shorter than that for males. The reason is believed to live in that the pitch duration of females is shorted than that of males, so the window size for females should be shorter than that of males to make the DFT analysis more accurate in the MFCC calculation. Compared with the GI feature extraction, this GD feature extraction can achieve better results, where the relative EER reductions are 21.7% and 12.2% for females and males, respectively. Experiments also show that the gender-matched cases outperform the gender-mismatched cases quite a lot. We believe it is due to the acoustical characteristics differences between females and males, which lead to the spatial distribution difference between female-specific UBM and male-specific UBM.

Future work will involve investigation on the generalization of experimental conclusions and exploration the relationship between the frame length and the step size. Also, we are looking forward to upgrading the gender-dependent issue to the pitch-dependent issue.
5. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No.61271389/61371136 and the National Basic Research Program (973 Program) of China under Grant No.2013CB329302.

6. REFERENCES


