Pitch-based Gender Identification with Two-stage Classification

Yakun Hu, Dapeng Wu, and Antonio Nucci

Abstract

In this paper, we address the speech-based gender identification problem. Mel-Frequency Cepstral Coefficients (MFCC) of voice samples are typically used as the features for gender identification. However, MFCC-based classification incurs high complexity. This paper proposes a novel pitch-based gender identification system with a two-stage classifier to ensure accurate identification and low complexity. The first stage of the classifier identifies and labels all the speakers whose pitch clearly indicates the gender of the speaker; the complexity of this stage is very low since only threshold-based decision rule on a scalar (i.e., pitch) is used. The ambiguous voice samples from all the other speakers (which cannot be classified with high accuracy by the first stage, and can be regarded as suspicious speakers or difficult cases) are forwarded to the second-stage for finer examination; the second-stage of our classifier uses Gaussian Mixture Model (GMM) to accurately isolate voice samples based on gender. Experiment results show that our system is speech language/content independent, microphone independent, and robust against noisy recording conditions. Our system is extremely accurate with probability of correct classification of 98.65%, and very efficient with about 5 seconds required for feature extraction and classification.

Index Terms

Gender Identification, Pitch, Energy Separation, Suspicious Speaker Detection, Gaussian Mixture Model (GMM)

I. Introduction

Gender identification is an important step in speaker and speech recognition systems [1], [2]. In both 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. [3], [4]. In content based multimedia indexing, speaker's gender is a cue used in the annotation. Thus, automatic gender identification is an important tool in multimedia signal analysis systems [5]–[7].

For speech signal based gender identification, the most commonly used features are pitch period and Mel-Frequency Cepstral Coefficients (MFCC) [7]. 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 [8]. 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 [9]–[11]. Second, an overlap of the pitch values between male and female voices naturally exists as shown in Fig.1. [7], thus making it a non-trivial problem to solve.

Yakun Hu and Dapeng Wu are with Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611. Correspondence author: Prof. Dapeng Wu, wu@ece.ufl.edu, http://www.wu.ece.ufl.edu. Antonio Nucci is with Narus, Inc., 570 Maude Court, Sunnyvale, CA 94085.

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 [12], the authors decide 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), thus 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 [5], [13], [14]. 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 pitch period but overcomes the limitations of pitch-based gender identification systems. We estimate the pitch period of a speech sample as sums of amplitude modulation-frequency modulation (AM-FM) formant models. AM components represent the envelope of the short-time speech signals which only contains information within a certain bandwidth, hence the noise effect is less severe. Since the possible distortion caused by the change of the recording may only occur at a certain bandwidth, the distortion effect becomes less severe too. For this reason the influence of language, microphone, and noise are much reduced in our gender identification system.

As mentioned earlier, a drawback of the pitch period feature is the accuracy of the final classification. In our system, we address this by using two (or more) steps in the classification stage. The first stage identifies and classifies all the speakers whose speech samples are unambiguous, i.e., these speakers can be classified as male or female without any doubt. The second stage operates on only those users whose voice sample could not be classified in the first stage. We call these speakers as “suspicious speakers” and use Gaussian Mixture Model (GMM) classifier to classify them. Our experimental results show that our system can achieve over 98% accuracy with very small computational overhead compared to the existing techniques. We also find that our system is robust to background noise, microphone variations, and language spoken by the speaker.

The rest of the paper is organized as follows. In section II, we present the architecture of our gender identification system. We discuss the pitch period estimation in Section III and describe the two-stage

classifier in Section IV. In Section V, we demonstrate the accuracy and efficiency of our system and conclude the paper in Section VI.

II. System Architecture

The architecture of the proposed gender identification system is shown in Fig. 2. For every speaker, a set of pitch period estimations are obtained from his or her speech signal. All pitch period estimations form a feature vector which is fed into the following classifier. Then, the gender decision of the speaker is made.

Fig. 3 describes the process of how to estimate the pitch period from the speech signal. For a speech signal, several vowel-like frames are first extracted. Then we obtain formant estimations for these vowel-like frames respectively. By bandpass filtering with the formant frequency as the center frequency, the corresponding vowel-like frames are bandpass filtered. The energy separation algorithm is then applied to the filtered frames and the AM components and FM components are separated. The last step is to estimate the periods of the quasi-periodic AM components and take them to be the pitch period estimations. All the estimations obtained from different frames form a vector of pitch values. In the figure, the multiple parallel arrows between the two consecutive blocks represent multiple frames, multiple components and multiple corresponding estimations

The overall structure of the classifier is shown in Fig. 4. The pitch feature vectors are fed into the first-stage classifier. In this stage, a simple thresholding method is applied to give a quick gender identification. For those speakers whose pitch values do not fall in the overlap of pitch values between male speakers and female speakers, gender decisions can be safely made. Those speakers are declared as the so-called unsuspicious speakers. For the other speakers whose pitch values fall in the overlap of pitch values between male speakers and female speakers, gender decisions are not able to be made by simple thresholding classifier and they will be declared as the so-called suspicious speakers. All suspicious ones will be further classified by the second-stage classifier using GMM method. The whole process is just like the normal check-in process in the airport. The ordinary people are checked in a very quick way while some suspicious ones need a careful inspection. By the two-stage classifier, the gender of all speakers can be identified correctly.

III. Pitch Feature Extraction

In this section, we describe how to accurately extract the pitch feature for gender identification. The detailed process of pitch period estimation from the speech signal is shown in Fig. 5. This method is based on AM-FM formant models of the speech signal and the energy separation algorithm which is able

to separate the AM components and the FM components. Then the pitch feature is obtained by estimating the period of the quasi-periodic AM component. All important components of the method are specifically described as follows.

A. AM-FM Formant Models

There are several evidences for the existence of modulations in speech signal [15]. From the theoretical point of view, during speech production, the air jet flowing through the vocal tract is highly unstable and oscillates between its walls. Hence, it changes the effective cross-sectional areas and air masses and affects the frequency of a speech resonance. Meanwhile, vortices can easily build up and encircle the air jet passing through. These vortices can act as modulators of the air jet energy. Moreover, it is well known that slow time variations of the oscillator elements can result in amplitude or frequency modulation. Thus, during speech production, the time-varying air masses and effective cross sectional areas of vocal tract cavities that rapidly vary following the separated airflow can cause modulations. Also from experiment, if the energy operator is applied in the bandpass filtered speech vowel signals around their formants, several pulses are often yielded. These energy pulses indicate some kind of modulation in each formant. Due to the description above, we incline to model speech signals using AM-FM formant models.

The AM-FM formant model has been successfully applied for speech analysis and modeling [16], speech synthesis, speech recognition and speaker identification. It is a nonlinear model that describes a speech resonance as a signal with a combined AM and FM structure:

$$r(t) = a(t)cos(\omega_c t + \omega_m \int_0^t q(\tau)d\tau + \theta) \tag{1}$$

where ω_c is the "center value" of the formant frequency, $q(t)$ is the frequency modulating signal, and $a(t)$ is the time-varying amplitude. The instantaneous formant frequency signal is defined as

$$\omega_i(t) = \omega_c + \omega_m q(t) \tag{2}$$

Usually we have $-1 < q(t) < 1$ and then ω_m characterizes the deviation of the instantaneous formant frequency around its "center value" and it denotes the maximum shift away from the "center value" ω_c.

The total short-time speech signal $s(t)$ is modeled as sums of K such AM-FM signals, one for each formant

$$s(t) = \sum_{k=1}^{K} r_k(t) \tag{3}$$

B. Energy based Speech Frame Extraction

In the practical system, what we are working with is fluent speech. We will extract the speech frames which contain relatively more energy from the influent speech and estimate the pitch values from these speech frames. Here, the shot-time analysis interval extracted from the long-time fluent speech wave is called a frame. Frames that contain relatively more energy can be determined according to different situations, i.e. top 10 frames containing the most energy among all, top 10% frames containing the most energy among all, etc. There are many reasons for using only the frames containing relatively more energy. On one hand, speech frames which contain relatively more energy could provide stronger pitch feature for gender identification. On the other hand, to decompose speech into AM components and FM components, we need to do formant estimation and extract every AM-FM resonance corresponding to each formant by bandpass filtering the speech signal around all its formants. [17] indicated that the acoustic characteristics of the obstruent sounds are not well represented through formants and the spectral characteristics of the noise source tend to mask the vocal tract resonances. Thus, the formants tracking are only suitable to the sonorant speech. Furthermore, stable pitch features should be obtained from voiced sounds. Voiced and sonorant speech frames usually contain relatively more energy.

In practice, for a given fluent speech signal, we extract speech frames by continually shifting a window over the speech signal. The length of the window is termed as the frame length and the window shifting interval is termed as the frame interval. In our system, the frame length is 2048 samples. With 2048 samples, the resolution of the estimated fundamental frequency (reciprocal of pitch period) can reach to about 10 Hz. This resolution value proved to be able to achieve a good gender identification performance. The frame interval is set as about 20 samples. The energy of each frame is calculated by

$$E(\vec{s_i}) = \sum_{n=1}^{l} s_i^2(n) \tag{4}$$

where $\vec{s_i} = [s_i(1), s_i(2), \cdots, s_i(n)]$ denotes the ith frame extracted from the fluent speech signal and l is the frame length. Energy of all the frames are ordered and the top ones are selected for the following process to obtain the pitch feature. [18] determines the voiced frame and the sonorant frame by calculating the energy contained within certain bandwidths. In our system, we just simply calculate the energy by using (4). No doubt, 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.

C. Pre-emphasis and Windowing

After all frames which contain relatively more energy are obtained, we will use linear predictive coding (LPC) analysis to estimate formant frequencies. The use of pre-emphasis pre-filtering is suggested to condition the speech frame before any following analysis. There are several justifications for this operation [19]. From a theoretical point of view, a proper pre-filter may remove the effects of glottal wave shape and the radiation characteristics of the lip. This will leave the all-pole vocal tract filter for analysis without wasting the LPC poles on glottal and radiation shaping. From a spectrum point of view, any preliminary flattening of the overall input spectrum before LPC processing allows the LPC analysis to do its own job of spectrum flattening better. Basically, these two statements imply that proper speech pre-emphasis will reduce the order of an LPC fit needed to do an equivalent spectrum match. Finally, from the point of view of a finite word length implementation, the proper pre-emphasis will reduce numerical error.

The speech signal pre-emphasis is performed by calculating its first-order difference. The new filtered speech signal is given by

$$\hat{s}_i(n) = s_i(n) + a * s_i(n-1) \tag{5}$$

where $s_i(n)$ is the input speech signal and $\hat{s}_i(n)$ is the pre-emphasis filtered speech signal. An optimal value for a can be obtained by solving for the filter that makes $\hat{s}_i(n)$ "white". This is given by the first order predictor, where

$$a = -\frac{R(1)}{R(0)} \tag{6}$$

$R(1)$ and $R(0)$ are autocorrelation coefficients of the input speech signal. The filtered signal is then guaranteed to have a smaller spectral dynamic range.

In order to extract a short-time interval from the pre-emphasis filtered speech signal for calculating the autocorrelation function and spectrum, the pre-emphasis filtered speech signal must be multiplied by an appropriate time window. The multiplication of the speech signal by the window function has two effects [20]. First, it gradually attenuates the amplitude at both ends of the extraction interval to prevent an abrupt change at the endpoints. Second, the multiplication of the speech frame by an appropriate window reduces the spectral fluctuation due to the variation of the pitch excitation position within the analysis interval. This is effective in producing stable spectra. As the windowing produces the convolution for the Fourier transform of the window function and the speech spectrum, or the weighted moving average in the spectral domain, it is thus desirable that the window function satisfy two characteristics in order to reduce the spectral distortion caused by the windowing. One is a high-frequency resolution, principally,

a narrow and sharp main lobe. The other is a small spectral leak from other spectral elements produced by the convolution, in other words, a large attenuation of the side lode. In practice, Hamming window, Hanning window, etc. are often used. In our system, Hamming window is adopted.

D. Formant Estimation

After the pre-emphasis and windowing, next we need to do formant estimation. Formant frequency is one of the most useful speech parameters which is specified by a vocal tract shape or its movements in various pronunciations. As mentioned in section III-A, the total short-time speech signal is modeled as the sums of K such AM-FM signals, one for each formant. Thus, accurate formant estimation is very important for extracting all AM-FM resonances.

The formants are physically defined as poles in a system function expressing the characteristics of a vocal tract. However, capturing and tracking formants accurately from natural speech is not so easy because of the variety of speech sounds. The frequencies at which the formants occur are primarily dependent upon the shape of the vocal tract, which is determined by the positions of the articulators (tongue, lips, jaw. etc.). In continuous speech, the formant frequencies vary in time as the articulators change position. The two historically representative methods for estimating formant frequencies are the analysis-by-synthesis (A-b-S) method and the LPC method [21]. The ideas are brilliant and many modified methods have stemmed from them [22], [23]. All these methods are ultimately based on the best matching between a spectrum to be analyzed and a synthesized one so that formant frequencies are estimated through spectral shapes. Hence, the estimation may be sensitive to spectral distortion and modifications.

In our system, after preprocessing by pre-emphasis and windowing, the speech frame is first separated into 4 shorter segmentations, each of which has 512 samples. Each segmentation with 512 samples is considered to be stationary. Thus, the linear prediction analysis can be applied for each segmentation to obtain the linear prediction coefficients that optimally characterize its short-time power spectrum. Generally, the power spectral shape has a smaller change within such a shorter interval, hence the LPC analysis of these shorter segmentations should be more robust to spectral distortion and modifications. Root-finding algorithm is then employed to find the zeros of the LPC polynomial. The zeros correspond to peaks in the short-time power spectrum and thus indicate the locations of the formant frequencies. The transformation from complex root pairs $z = re^{\pm j\theta}$ and sampling frequency f_s to formant frequency F and 3-dB bandwidth B are as follows [24]:

$$F = \frac{f_s}{2\pi}\theta Hz \tag{7}$$

$$B = -\frac{f_s}{\pi} lnr \tag{8}$$

The order selection of the LPC model is important to accurate formant estimation. If the order is chosen smaller, the short-time power spectrum can't be fully characterized and it may lead to missing peaks. If chosen larger, the speech signal is over-determinedly modeled and spurious peak may occur. In our experiment, the order for the analysis is set to be 13. It seems to be a good choice which can yield satisfactory formant estimation.

For each segmentation, as more than one zeros of the LPC polynomial can be found, more than one formant frequencies are obtained. We select the minimum one which contains the most speech energy. Then for each frame, four estimations of the formant frequency are obtained. Generally, the four estimations are close to each other and all of them contains the most speech energy of each segmentation. Among the four, we again select the minimum one as the final formant estimation for the frame. This method is proved to be able to yield a good formant estimation with a relatively low computation complexity.

The formant estimation is then used as the center frequency to bandpass filter the corresponding speech frame. Gabor filter is used as the bandpass filter, whose impulse and frequency responses are

$$h(t) = exp(-\alpha^2 t^2) cos(\omega_c t) \tag{9}$$

$$H(\omega) = \frac{\sqrt{\pi}}{2\alpha}(exp[-\frac{(\omega-\omega_c)^2}{4\alpha^2}] + exp[-\frac{(\omega+\omega_c)^2}{4\alpha^2}]) \tag{10}$$

where ω_c is the center value of the formant frequencies obtained above. The reasons for selecting the above bandpass filter are twofold: $1)$ It is optimally compact in the time and frequency width product assumes the minimum value in the uncertainty principle inequality; $2)$ The Gaussian shape of $H(\omega)$ avoids producing side-lobes (or big side-lobes after truncation of h) that could produce false pulses in the output of the latter energy separation.

Here, a problem could be how to determine the bandwidth of the Gabor filter when doing the bandpass filtering. The 3-dB bandwidth of the Gabor filter is equal to $\alpha/\sqrt{2\pi}$. The bandwidth should not be too wide because then they will include significant contributions from neighbouring formants which may cause parasitic modulations. On the other hand, the Gabor filters should not have a very narrow bandwidth because this would miss or deemphasize some of the modulations. In our system, a 3-dB bandwidth of 400Hz is used. Experimental results indicate that it could be a suitable choice.

E. Energy Separation

All the corresponding bandpass filtered frames are obtained and the AM components and FM components needs to be decomposed. We use the "energy-tracking" operator to estimate the amplitude envelope

$|a(t)|$ and the instantaneous frequency $\omega_i(t)$ [15].

For continuous-time signal, the energy operator is defined as

$$\psi_c[x(t)] = [\dot{x(t)}]^2 - x(t)\ddot{x(t)} \tag{11}$$

where $\dot{x(t)} = dx/dt$ and $\ddot{x(t)} = d\dot{x(t)}/dt$. For discrete-time signal, the energy operator is defined as

$$\psi_d[x(n)] = x(n)^2 - x(n-1)x(n+1) \tag{12}$$

where $n = 0, \pm 1, \pm 2, \cdots$ It can be concluded from [15] that for any constants A and ω_c, we have

$$\psi_c[Acos(\omega_c t + \theta)] = (A\omega_c)^2 \tag{13}$$

For time-varying amplitude and frequency, [25] shows that

$$\psi_c[a(t)cos(\int_0^t \omega_i(\tau)d\tau_\theta)] = (a(t)\omega_i(t))^2 \tag{14}$$

Assuming that the signals $a(t)$ and $\omega_i(t)$ do not vary too fast or too greatly in time compared to ω_c. Thus, the combined use of the energy operator on the AM-FM signal and its derivative (or difference) can lead to an elegant algorithm for separately estimating the amplitude signals and the frequency signals. In our experiments, the discrete-time signal is considered. The discrete energy separation algorithm (DESA) is shown as follows:

$$x(n) - x(n-1) = y(n) \tag{15}$$

$$arccos(1 - \frac{\psi[y(n)] + \psi[y(n+1)]}{4\psi[x(n)]}) \approx \omega_i(n) \tag{16}$$

$$\sqrt{\frac{\psi[x(n)]}{1 - (1 - \frac{\psi[y(n)]+\psi[y(n+1)]}{4\psi[x(n)]})^2}} \approx |a(n)| \tag{17}$$

It is very simple to implement DESA since it only requires a few simple operations per output sample and involves a very short window of samples around the time instant at which we estimate the amplitude and frequency.

F. Pitch Period Estimation

The amplitude envelope $a(n)$ obtained by DESA is a quasi-periodic signal. Actually, its period is a good estimation of the pitch period. By estimating the period of $a(n)$, we can obtain the pitch period.

The formant frequency mainly depends on the vocal tract shape and the positions of the articulators (tongue, lips, jaw, etc.). It must be different in various pronunciations even for the same speaker. Thus, the formant frequency is a content-dependent feature and not a stable feature for gender identification. Pitch

represents the perceived fundamental frequency of a sound. Usually male speakers have relatively lower fundamental frequency values while the female speakers have relatively higher fundamental frequency values. Also it is relatively stable for a specific speaker. Thus, it could be a good feature for gender identification.

Power spectrum analysis is used to estimate the pitch period. The quasi-periodicity of the amplitude envelope in the time domain would yield peaks in the corresponding power spectrum. Thus, the problem of pitch period estimation can be converted into the peak detection in the power spectrum. In the power spectrum of the amplitude envelope, we search for the largest non-dc peak and take the reciprocal of its frequency location as our estimation of the pitch period. The resolution of the fundamental frequency (reciprocal of pitch period) can reach to about 10 Hz. To increase the resolution of the estimation, the frame length needs to be increased. As the formant frequencies may have considerable change within a longer interval, the formant estimation and hence the pitch period estimation may not be accurate enough by using a longer frame. A frame length with 2048 samples seems to make a good tradeoff between the accuracy and the resolution. 10 Hz resolution proves suitable for accurate gender identification.

IV. Two-Stage Classifier with Suspicious Speaker Detection

Section III specified how to obtain the pitch feature from the speech signal. Now the pitch feature will be fed into the classifier to make gender decisions for speakers. Section II roughly describes the structure of the two-stage classifier. The detailed structures of the proposed two-stage classifier with suspicious speaker detection scheme during the training phase and the testing phase are shown in Fig. 6 and Fig. 7. In the training phase, we put all the vectors of the pitch feature of all speakers into a matrix $P_{i,j}$, where i denotes the pitch index and j denotes the speaker index. For the kth column vector, i.e. the vector of the pitch feature of speaker k, the most frequent pitch value P_k is extracted. Based on the most frequent pitch values of all speakers, two thresholds P_M and P_F are set to make sure that all speakers whose most frequent pitch values are smaller than P_M are male and all speakers whose most frequent pitch values are larger than P_F are female. The rest speakers are thought to be suspicious speakers who need to be further processed by the second-stage classifier. That is to say, if $P_k < P_M$, the speaker k must be a male, if $P_k > P_F$, then the speaker k must be a female, if $P_M \leq P_k \leq P_F$, then speaker k is declared as suspicious speaker. Suppose $\bar{P}_M$ is the vector of the most frequent pitch values of all male speakers and $\bar{P}_F$ is the vector of the most frequent pitch values of all female speakers. A simple method to determine the two thresholds is to let $P_M = min\bar{P}_F$ and let $P_F = max\bar{P}_M$. Here we have $P_M < P_F$ because of the pitch value overlap between the male speakers and female speakers. By this threshold setting, we are able

to ensure that all speakers can be grouped into male speaker cluster, female speaker cluster and suspicious speaker cluster at the first stage of the gender identification. Actually, the threshold can be determined in a more general way: $P_M \leq min\bar{P}_F$ and $P_F \geq max\bar{P}_M$. The larger the interval of the two thresholds, the more speakers will be declared as suspicious speakers and more reliable the gender identification will be at the first stage classifier. However, of course, more work needs to be done in the second-stage gender identification. The total computation complexity is increased. Thus, the thresholds should be set according to the requirement of the practical application. Two more things should be further pointed out. One is the resolution of the pitch values. As the most frequent pitch values of all speakers is used to set the thresholds, the resolution of the pitch values should be carefully chosen. If too large, the gender identification performance may not be good enough. If too small, the most frequent pitch values may not well represent all pitch period estimations. In our experiments, the 10Hz resolution proves to be a good choice. The other thing is that for one speaker, sometimes there are more than one most frequent pitch values, i.e. more than one pitch values occur with equal highest frequency. Under this condition, the most frequent pitch value of this speaker is determined in this way: the pitch value occurs with the highest frequency and being closest to the mean value of all pitch values is considered as the most frequent pitch value of this speaker.

At the second-stage gender identification for suspicious speakers, GMM method is applied. Both GMMs of male speakers and female speakers are trained by Expectation Maximization (EM) algorithm, using the pitch feature vectors of all male speakers and all female speakers,respectively. Both GMMs are initialized by k-mean clustering. The dimension of the pitch feature vectors used for training is adjustable. The vector of the pitch values obtained for each speaker can be segmented into several lower-dimension feature vectors. These lower-dimension feature vectors can be used for training. The lower the dimension is, the more training samples are available. Coupled with the feature dimension, the order of GMMs is another adjustable parameter which associates with the computation complexity and the gender identification performance.

During the testing phase, for every speaker, e.g. speaker k, we compare his or her most frequent pitch value P_k with the thresholds P_M and P_F determined in the training phase. If $P_k < P_M$, speaker k is classified as male speaker. If $P_k > P_F$, speaker k is classified as female speaker. If $P_M \leq P_k \leq P_F$, speaker k is classified as suspicious speaker. For each suspicious speaker, we feed the feature vectors of his or her pitch values (with the same dimension used in the training phase) into GMMs of male speakers and female speakers, respectively. Suppose the feature vector is denoted by $v_{i,j}$ where $i = 1, 2, \cdots$ denotes the feature vector index and j denotes the speaker index. Also the GMMs of male speakers and female

speakers are denoted by f_M and f_F. Thus, the output of two GMMs are obtained by $\sum_i log(f_M(v_{i,j}))$ and $\sum_i log(f_F(v_{i,j}))$. All feature vectors contribute to the GMM output. We select the one which has the larger output. If the GMM of male speakers yields a larger output than the GMM of the female speakers, then the suspicious speaker is classified as male speaker. Otherwise, the suspicious speaker is classified as female speaker. From the description above, we can know that the whole classifier consists of two stages which are separately quick stage using simple thresholding and slow stage using GMM. The advantages of completing the gender identification in two stages include the computation complexity reduction and performance improvement. In the aspect of the computation complexity, as we always use the simpler method first to do the gender identification, the computation complexity are reduced at the largest extent. However, using simple methods for gender identification, the performance may not be reliable. That is the reason why we pick the suspicious speakers out and use more complicated methods to ensure the excellent gender identification performance. The two-stages gender identification can be extended to the multi-stage gender identificaiton till the gender identification results of all speakers are believed to be reliable and no speaker is declared as suspicious speaker.

V. Experimental Results

Experiments are carried out to validate the excellent performance of the gender identification system proposed in this paper. Also the experimental results are shown to validate the language independence, microphone independence and robustness to the noise condition of our proposed gender identification system. In our experiments, the TIDIGITS dataset is used. Also we recorded speech for several male speakers and female speakers to help carry out our experiments.

A. Gender Identification on TIDIGITS

To test the performance of our proposed gender identification system, the experiment is carried out on TIDIGITS dataset.

In our experiment, read utterances from 111 men and 111 women in TIDIGITS dataset are used. 77 sequences of these digits were collected from each speaker. The data were collected in a quiet environment with the microphone placed 2-4 inches in front of the speaker's mouth and digitized at 20 kHz. For the 77 sequences from each speaker, 39 sequences are used for training and the rest 38 sequences are used for testing. For every sequence, only the speech frame which has the largest energy is extracted and the pitch period is estimated from that frame. Thus, for each speaker, 39 pitch values are estimated for training and 38 pitch values are estimated for testing.

TABLE I
COMPARISON OF CLASSIFIERS

Classifier	Identification Rate	Time	Data needed to be stored in memory
Pitch Thresholding + GMM	98.65%	5.6078s	Pitch Values of all Suspicious Speakers, GMM Parameters
Pitch Thresholding	96.85%	5.4848s	Most Frequent Pitch Values of all Speakers
GMM	98.2%	5.6217s	Pitch Values of all Speakers, GMM Parameters

For the pitch feature extraction process, the experiment shows that for 111 male speakers and 111 female speakers, a total of 1217.5s is spent. That is to say, for every speaker, about 5.5s is needed for the pitch feature extraction. This is fast enough for the real-time application of our proposed system.

For the gender identification process, training and testing are separately carried out. In the training phase, among the most frequent pitch values of all male speakers,the maximum value is 185.55Hz. Among the most frequent pitch values of all female speakers,the minimum value is 156.3 Hz. Thus, the thresholds can be set as 156.3 Hz and 185.55 Hz. All the speakers whose most frequent pitch values fall between 156.3 Hz and 185.55 Hz are declared as suspicious speakers. For suspicious speakers, the second stage GMM classifier is applied. GMMs of male speakers and female speakers are trained by 2-dimension pitch feature vectors of all male speakers and all female speakers, respectively. In the training phase, there are 39 pitch values for each speaker. Thus, for GMMs of both male speakers and female speakers, $19 \times 111 = 2109$ pitch feature vectors are available for training. The orders of both GMMs are set as 5 and both GMMs are initialized by k-means clustering. In the testing phase, if the most frequent pitch value of a speaker is larger than 185.55Hz, then this speaker is declared as a female speaker. If the most frequent pitch value of a speaker is smaller than 156.3 Hz, then this speaker is declared as a male speaker. Otherwise, this speaker is declared as a suspicious speaker and needs to be classified in the second stage by using GMMs. In the first stage gender identification using simply the thresholding method, 10 out of 111 male speakers are declared as suspicious speakers and 14 out of 111 female speakers are declared as suspicious speakers. For each suspicious speaker, the 2-dimension pitch feature vector of his or her pitch values for testing are fed into both GMMs. In the testing phase, there are 38 pitch values for each speaker. Thus, for each speaker, totally $19 \times 111 = 2109$ pitch feature vectors are available for testing. The model who yield the larger output will be selected. The output calculation has been described in IV. In this way, the gender decision of every suspicious speaker is made. Table I summarizes our two-stage classifier in the aspect of gender identification performance and computation complexity (measure in time cost of gender identification for each speaker) and make a comparison among our classifier, the pitch thresholding classifier and GMM classifier.

From Table I, we know that our proposed two-stage classifier can achieve 98.65% correct gender

identification rate which is nearly the same as the GMM classifier (98.2%) but is better than the pitch thresholding classifier (96.85%). On the other hand, to compare the time load and the memory load of the proposed two-stage classifier and the GMM classifier both of which achieve the excellent performance, the proposed two-stage classifier spends less time to complete the gender identification for all speakers and requires less memory than the GMM classifier.

According to [14], the classifier combining pitch and MFCC usually achieves about 98% correct gender identification rate. However, MFCC calculation requires much more computation and memory. Also MFCC has the problem of over-training. It learns too much detail from the speech signal. Thus, It is not a good feature for gender identification. Although it can achieve good performance in the perfect recording condition (i.e. no noise distortion, no microphone change, etc.), it is not able to work in the varying recording condition. Compared with it, our proposed system only uses pitch feature and adopts two-stage gender identification with suspicious speaker detection scheme to reduce the computation complexity and memory requirement while maintaining a good performance. Thus, our proposed system has great advantage not only in identification performance but also in the computation complexity and memory requirement. Furthermore, our proposed system is believed to be able to work well in the varying recording condition. The following discussion will show that our proposed system have the characteristics of language independence, microphone independence and being robust to the background noise and strong additive white Gaussian noise (AWGN).

B. Language Independence and Content Independence

Speakers are from all different countries and speak different kinds of language. A good gender identification system should be robust to all speakers speaking different kinds of language and content, i.e. a good gender identification system should be language/content independent. This experiment is carried out to study whether our proposed system possesses the characteristics of language/content independence or not.

In our experiment, a one-minute clean Mandarin fluent speech and a one-minute clean English fluent speech are respectively recorded for male speaker A and female speaker B with the same microphone in a quiet environment. The sampling frequency is 22050 Hz and the number of bits per sample to encode the data is 16. The pitch feature is extracted in the way described in section III. 40 pitch values are estimated for every fluent speech uttered by the speakers.

Table II and Table III separately summarize the result for the male speaker and the female speaker by listing the most frequent value (mode value), the mean value and the standard deviation of every pitch

TABLE II
PITCH PERIOD ESTIMATIONS OF MALE SPEAKER A WITH DIFFERENT KINDS OF LANGUAGE (IN HZ)

Language	Mode Value	Mean Value	Standard Deviation
English Speech	139.9658 Hz	139.9658 Hz	0 Hz
Mandarin Speech	129.1992 Hz	129.1992 Hz	0 Hz

TABLE III
PITCH PERIOD ESTIMATIONS OF FEMALE SPEAKER B WITH DIFFERENT KINDS OF LANGUAGE (IN HZ)

Language	Mode Value	Mean Value	Standard Deviation
English Speech	247.6318 Hz	241.4410 Hz	16.6951 Hz
Mandarin Speech	269.1650 Hz	275.8942 Hz	10.8182 Hz

feature vector which consists of 40 estimations.

From Table II and Table III, we know that for both male speaker A and female speaker B, no matter they use English or Mandarin, the pitch feature extracted from the speech signals is pretty stable. The standard deviation listed above approximates to the estimation resolution of the pitch values which is about 10 Hz. Even with some kind of deviation, the mode value and the mean value still fall in the interval of his or her gender category. Hence, it will not affect the final result of their gender identification. From this point of view, we are able to say that our proposed system exhibits some characteristics of language/content independence. In fact, all languages share many common phonemes. No matter what language a speaker speaks, the fundamental frequency (i.e. reciprocal of pitch period) of his or her voice does not change. As the pitch period estimation method is speech content independent, if the pitch period estimation method can work well, it should be language/content independent.

C. Microphone Independence

In practice, speakers do not always use the same microphone to record their speech. A good gender identification system should be robust to microphone change during the training phase and the testing phase. That is, a good gender identification system should be microphone independent. This experiment is carried out to study whether our proposed system possesses the characteristics of microphone independence or not.

In our experiment, two one-minute clean English fluent speech are respectively recorded for male speaker C and female speaker D with two different microphones in a quiet environment. The sampling frequency is 22050 Hz and the number of bits per sample to encode the data is 16. The pitch feature is extracted in the way described in section III. 40 pitch values are estimated for every fluent speech uttered by the speakers.

TABLE IV
PITCH PERIOD ESTIMATIONS OF MALE SPEAKER C WITH DIFFERENT MICROPHONES (IN HZ)

Microphone	Mode Value	Mean Value	Standard Deviation
Mic 1	129.1992 Hz	130.0067 Hz	5.6592 Hz
Mic 2	129.1992 Hz	130.0067 Hz	5.6592 Hz

TABLE V
PITCH PERIOD ESTIMATIONS OF FEMALE SPEAKER D WITH DIFFERENT MICROPHONES (IN HZ)

Microphone	Mode Value	Mean Value	Standard Deviation
Mic 1	247.6318 Hz	241.4410 Hz	16.6951 Hz
Mic 2	258.3984 Hz	268.0884 Hz	11.3839 Hz

Table IV and Table V separately summarize the result for the male speaker and the female speaker by listing the most frequent value (mode value), the mean value and the standard deviation of every pitch feature vector which consists of 40 estimations.

From Table IV and Table V, we know that for both male speaker C and female speaker D, no matter what microphone they use, the pitch feature extracted from the speech signals is pretty stable. The standard deviation listed above approximates to the estimation resolution of the pitch values which is about 10 Hz. Even with some kind of deviation, the mode value and the mean value still fall in the interval of his or her gender category. Hence, it will not affect the final result of their gender identification. From this point of view, we are able to say that our proposed system exhibits some characteristics of microphone independence.

To further validate the microphone independence of our proposed system, another experiment is carried out.

We recorded speech for 3 male speakers and 3 female speakers with two different microphones. We use all speech recorded by one microphone for training and use all speech recorded by another microphone for testing.

In the training phase, the thresholds are determined as $P_M = 183.0322Hz$ and $P_F = 226.0986Hz$. In the testing phase, the most frequent pitch values of the 3 male speakers are 107.6660 Hz, 150.7324 Hz, 129.1992 Hz and the most frequent pitch values of the 3 male speakers are 215.3320 Hz, 269.1650 Hz, 290.6982 Hz. By using just the first-stage thresholding classifier, the correct gender identification rate reaches 100%.

Although the total number of speakers for the experiment are not very large, it did indicate the microphone independence of our proposed system. Actually, the microphone is like a filter. Different microphones lead to different filtering effect of the speech signal. Many existing methods suffer from

TABLE VI
PITCH PERIOD ESTIMATIONS OF MALE SPEAKER E IN SCENARIO OF AIR-CONDITIONER NOISE WITH DIFFERENT MICROPHONES AND DIFFERENT KINDS OF LANGUAGE (IN HZ)

Microphone/Language	Mode Value	Mean Value	Standard Deviation
Mic 1+English	172.2656 Hz	176.8414 Hz	5.3902 Hz
Mic 1+Mandarin	172.2656 Hz	171.7273 Hz	4.8457 Hz
Mic 2+English	183.0322 Hz	181.9555 Hz	7.2327 Hz

the microphone change. MFCC fails to work when microphone changes during the training phase and testing phase even for the case of very few speakers. As MFCC learns too much detail of the speech spectrum, it depends greatly on the recording condition such as the microphone condition. The microphone independence of our proposed system is a big advantage in the practical application.

D. Noise Independence

Sometimes the speech is recorded in an environment with background noise such as the air-conditioner noise, the background music noise, keyboard striking noise and road noise, etc. A good gender identification system should be robust to all kinds of background noise. The following experiments are carried out on several noise scenarios to study whether our proposed system is robust to the background noise and AWGN.

Case 1: Air-conditioner Noise

This experiment is carried out to study whether our proposed system possesses the characteristics of being robust to the air-conditioner noise or not and to study whether our proposed system possesses the characteristics of language independence and microphone independence in the scenario of air-conditioner noise or not.

In our experiment, for both male speaker E and female speaker F, two one-minute English fluent speech are recorded with two different microphones in the scenario of air-conditioner noise. Also a one-minute mandarin fluent speech is recorded with one of the two microphones but in the same scenario of air-conditioner noise. The air-condition noise hears pretty clear and can not be neglected. The sampling frequency is 22050 Hz and the number of bits per sample to encode the data is 16. The pitch feature is extracted in the way described in section III. 40 pitch values are estimated for every fluent speech uttered by the speakers.

Table VI and Table VII separately summarize the results for male speaker E and female speaker F by listing the most frequent value (mode value), the mean value and the standard deviation of every pitch feature vector which consists of 40 estimations.

TABLE VII
PITCH PERIOD ESTIMATIONS OF FEMALE SPEAKER F IN SCENARIO OF AIR-CONDITIONER NOISE WITH DIFFERENT MICROPHONES AND DIFFERENT KINDS OF LANGUAGE (IN HZ)

Microphone/Language	Mode Value	Mean Value	Standard Deviation
Mic 1+English	269.1650 Hz	265.9351 Hz	4.9967 Hz
Mic 1+Mandarin	269.1650 Hz	274.5483 Hz	5.4519 Hz
Mic 2+English	269.1650 Hz	272.9333 Hz	7.9193 Hz

From Table VI and Table VII, we know that for both male speaker E and female speaker F, even with different microphones and different kinds of language, in the scenario of air-conditioner noise which can not be neglected, the pitch feature extracted from the speech signals is pretty stable. The standard deviations listed above all have a smaller value than the estimation resolution of the pitch values which is about 10 Hz. Even with some kind of deviation, the mode value and the mean value still fall in the interval of his or her gender category. Hence, it will not affect the final result of their gender identification. From this point of view, we are able to say that our proposed system exhibits the characteristics of being robust to the air-conditioner noise and our proposed system exhibits the characteristics of microphone independence and language independence in the scenario of air-conditioner noise.

Case 2: Background Music Noise

This experiment is carried out to study whether our proposed system possesses the characteristics of being robust to the background music noise or not and to study whether our proposed system possesses the characteristics of language independence and microphone independence in the scenario of background music noise or not.

In our experiment, for both male speaker G and female speaker H, two one-minute English fluent speech are recorded with two different microphones in the scenario of background music noise. Also a one-minute mandarin fluent speech is recorded with one of the two microphones but in the same scenario of background music noise. The background music noise hears pretty clear and can not be neglected. The sampling frequency is 22050 Hz and the number of bits per sample to encode the data is 16. The pitch feature is extracted in the way described in section III. 40 pitch values are estimated for every fluent speech uttered by the speakers.

Table VIII and Table IX separately summarize the results for male speaker G and female speaker H by listing the most frequent value (mode value), the mean value and the standard deviation of every pitch feature vector which consists of 40 estimations.

From Table VIII and Table IX, we know that for both male speaker G and female speaker H, even with different microphones and different kinds of language, in the scenario of background music noise which

TABLE VIII
PITCH PERIOD ESTIMATIONS OF MALE SPEAKER G IN SCENARIO OF BACKGROUND MUSIC NOISE WITH DIFFERENT MICROPHONES AND DIFFERENT KINDS OF LANGUAGE (IN HZ)

Microphone/Language	Mode Value	Mean Value	Standard Deviation
Mic 1+English	161.4990 Hz	172.2656 Hz	10.9038 Hz
Mic 1+Mandarin	150.7324 Hz	155.5774 Hz	5.4246 Hz
Mic 2+English	139.9658 Hz	134.8517 Hz	5.4415 Hz

TABLE IX
PITCH PERIOD ESTIMATIONS OF FEMALE SPEAKER H IN SCENARIO OF BACKGROUND MUSIC NOISE WITH DIFFERENT MICROPHONES AND DIFFERENT KINDS OF LANGUAGE (IN HZ)

Microphone/Language	Mode Value	Mean Value	Standard Deviation
Mic 1+English	193.7988 Hz	191.9147 Hz	5.3902 Hz
Mic 1+Mandarin	193.7988 Hz	192.1838 Hz	5.7439 Hz
Mic 2+English	193.7988 Hz	209.6796 Hz	13.7896 Hz

can not be neglected, the pitch feature extracted from the speech signals is pretty stable. The standard deviations listed above are all less than or close to the estimation resolution of the pitch values which is about 10 Hz. Even with some kind of deviation, the mode value and the mean value still fall in the interval of his or her gender category. Hence, it will not affect the final result of their gender identification. From this point of view, we are able to say that our proposed system exhibits the characteristics of being robust to the background music noise and our proposed system exhibits the characteristics of microphone independence and language independence in the scenario of background music noise.

Case 3: Keyboard Striking Noise

This experiment is carried out to study whether our proposed system possesses the characteristics of being robust to the keyboard striking noise or not and to study whether our proposed system possesses the characteristics of language independence and microphone independence in the scenario of keyboard striking noise or not.

In our experiment, for both male speaker I and female speaker J, two one-minute English fluent speech are recorded with two different microphones in the scenario of keyboard striking noise. Also a one-minute mandarin fluent speech is recorded with one of the two microphones but in the same scenario of keyboard striking noise. The keyboard striking noise hears pretty clear and can not be neglected. The sampling frequency is 22050 Hz and the number of bits per sample to encode the data is 16. The pitch feature is extracted in the way described in section III. 40 pitch values are estimated for every fluent speech uttered by the speakers.

Table X and Table XI separately summarize the results for male speaker I and female speaker J by listing the most frequent value (mode value), the mean value and the standard deviation of every pitch

TABLE X
PITCH PERIOD ESTIMATIONS OF MALE SPEAKER I IN SCENARIO OF KEYBOARD STRIKING NOISE WITH DIFFERENT MICROPHONES AND DIFFERENT KINDS OF LANGUAGE (IN HZ)

Microphone/Language	Mode Value	Mean Value	Standard Deviation
Mic 1+English	161.4990 Hz	164.1907 Hz	6.3345 Hz
Mic 1+Mandarin	150.7324 Hz	153.6932 Hz	5.4451 Hz
Mic 2+English	150.7324 Hz	149.1174 Hz	3.8934 Hz

TABLE XI
PITCH PERIOD ESTIMATIONS OF FEMALE SPEAKER J IN SCENARIO OF KEYBOARD STRIKING NOISE WITH DIFFERENT MICROPHONES AND DIFFERENT KINDS OF LANGUAGE (IN HZ)

Microphone/Language	Mode Value	Mean Value	Standard Deviation
Mic 1+English	193.7988 Hz	193.7988 Hz	0 Hz
Mic 1+Mandarin	193.7988 Hz	193.7988 Hz	0 Hz
Mic 2+English	193.7988 Hz	199.1821 Hz	15.2263 Hz

feature vector which consists of 40 estimations.

From Table X and Table XI, we know that for both male speaker and female speaker, even with different microphones and different kinds of language, in the scenario of keyboard striking noise which can not be neglected, the pitch feature extracted from the speech signals is pretty stable. The standard deviations listed above are all less than or close to the estimation resolution of the pitch values which is about 10 Hz. Even with some kind of deviation, the mode value and the mean value still fall in the interval of his or her gender category. Hence, it will not affect the final result of their gender identification. From this point of view, we are able to say that our proposed system exhibits the characteristics of being robust to the keyboard striking noise and our proposed system exhibits the characteristics of microphone independence and language independence in the scenario of keyboard striking noise.

Case 4: Noise from Moving Cars

This experiment is carried out to study whether our proposed system possesses the characteristics of being robust to noise from moving cars and to study whether our proposed system possesses the characteristics of language independence and microphone independence in the scenario of noise from moving cars.

In our experiment, for both male speaker K and female speaker L, two one-minute English fluent speech are recorded with two different microphones in the scenario of noise from moving cars. Also a one-minute mandarin fluent speech is recorded with one of the two microphones but in the same scenario of noise from moving cars. The noise from moving cars can be heard pretty clear and can not be neglected. The sampling frequency is 22050 Hz and the number of bits per sample to encode the data is 16. The pitch feature is extracted in the way described in section III. 40 pitch values are estimated for every fluent

TABLE XII
PITCH PERIOD ESTIMATIONS OF MALE SPEAKER K IN SCENARIO OF NOISE FROM MOVING CARS WITH DIFFERENT MICROPHONES AND DIFFERENT KINDS OF LANGUAGE (IN HZ)

Microphone/Language	Mode Value	Mean Value	Standard Deviation
Mic 1+English	150.7324 Hz	146.1566 Hz	4.8072 Hz
Mic 1+Mandarin	161.4990 Hz	164.5944 Hz	7.3030 Hz
Mic 2+English	172.2656 Hz	172.2656 Hz	0 Hz

TABLE XIII
PITCH PERIOD ESTIMATIONS OF FEMALE SPEAKER L IN SCENARIO OF NOISE FROM MOVING CARS WITH DIFFERENT MICROPHONES AND DIFFERENT KINDS OF LANGUAGE (IN HZ)

Microphone/Language	Mode Value	Mean Value	Standard Deviation
Mic 1+English	193.7988 Hz	193.7988 Hz	0 Hz
Mic 1+Mandarin	215.3320 Hz	218.8312 Hz	5.1070 Hz
Mic 2+English	204.5654 Hz	204.5654 Hz	0 Hz

speech uttered by the speakers.

Table XII and Table XIII separately summarize the results for male speaker K and female speaker L by listing the most frequent value (mode value), the mean value and the standard deviation of every pitch feature vector which consists of 40 estimations.

From Table XII and Table XIII, we know that for both male speaker K and female speaker L, even with different microphones and different kinds of language, in the scenario of noise from moving cars which can not be neglected, the pitch feature extracted from the speech signals is pretty stable. The standard deviations listed above are all less than or close to the estimation resolution of the pitch values which is about 10 Hz. Even with some kind of deviation, the mode value and the mean value still fall in the interval of his or her gender category. Hence, it will not affect the final result of their gender identification. From this point of view, we can say that our proposed system exhibits the characteristics of being robust to noise from moving cars and our proposed system exhibits the characteristics of microphone independence and language independence in the scenario of noise from moving cars.

Case 5: AWGN

In practice, sometimes speech signal will be recorded in the transmitter side and then be transmitted through an AWGN channel. In this condition, the original clean speech is corrupted by AWGN. Hence, performance evaluation of a gender identification system in the scenario of AWGN is important. This experiment is carried out to study whether our proposed system is robust to AWGN or not.

The setup condition of the experiments is the same as mentioned in V-A, except that the white Gaussian noise is added to every clean read utterance from the TIDIGITS dataset. The signal-to-noise ratio (SNR) is adjusted to see its effect on the gender identification performance. 0dB case and 5dB case are tested,

respectively. For each case, we conduct two different experiments. One is that the clean training samples is used for training and the testing samples corrupted by AWGN is used for testing. The other one is that the training samples corrupted by AWGN is used for training and the testing samples corrupted by AWGN is used for testing. In all experiments, due to the randomness of AWGN, the extracted pitch feature may not be as stable as in the scenario of clean speech. To make sure that in the testing phase, genders of all so-called unsuspicious speakers are correctly identified by using the first-stage thresholding classifier, the thresholds are determined in a different way from the one in V-A. P_F is set as being larger than the maximum one of the most frequent pitch values of all male speakers. P_M is set as being smaller than the minimum one of the most frequent pitch values of all female speakers. Thus, the interval between the two thresholds becomes larger and more speakers are declared as suspicious speakers and more reliable the identification decision will be made by using the first-stage thresholding classifier.

For the 5dB case, in the first experiment, we use the clean training samples for training and use the corrupted testing samples for testing. Among the most frequent pitch values of all male speakers,the maximum value is 185.55Hz. Among the most frequent pitch values of all female speakers,the minimum value is 156.3Hz. The thresholds are set as 146.48 Hz and 195.31 Hz. All the speakers whose most frequent pitch values fall between 146.48 Hz and 195.31 Hz are declared as suspicious speakers. For suspicious speakers, the second-stage GMM classifier is applied. The final correct gender identification rate is 97.75%. There are 109 out of 111 male speakers and 108 out of 111 female speakers whose gender are correctly identified. In the second experiment, we use the corrupted training samples for training and use the corrupted testing samples for testing. Among the most frequent pitch values of all male speakers,the maximum value is 185.55Hz. Among the most frequent pitch values of all female speakers,the minimum value is 156.3Hz. The thresholds are set as 146.48 Hz and 195.31 Hz. All the speakers whose most frequent pitch values fall between 146.48 Hz and 195.31 Hz are declared as suspicious speakers. For suspicious speakers, the second-stage GMM classifier is applied. The final correct gender identification rate is 97.3%. There are 108 out of 111 male speakers and 108 out of 111 female speakers whose gender are correctly identified.

For the 0dB case, in the first experiment, we use the clean training samples for training and use the corrupted testing samples for testing. Among the most frequent pitch values of all male speakers,the maximum value is 185.55Hz. Among the most frequent pitch values of all female speakers,the minimum value is 156.3Hz. The thresholds are set as 146.48 Hz and 195.31 Hz. All the speakers whose most frequent pitch values fall between 146.48 Hz and 195.31 Hz are declared as suspicious speakers. For suspicious speakers, the second-stage GMM classifier is applied. The final correct gender identification

rate is 96.85%. There are 110 out of 111 male speakers and 105 out of 111 female speakers whose gender are correctly identified. In the second experiment, we use the corrupted training samples for training and use the corrupted testing samples for testing. Among the most frequent pitch values of all male speakers,the maximum value is 175.78 Hz. Among the most frequent pitch values of all female speakers,the minimum value is 156.3Hz. The thresholds are set as 146.48 Hz and 185.55 Hz. All the speakers whose most frequent pitch values fall between 146.48 Hz and 185.55 Hz are declared as suspicious speakers. For suspicious speakers, the second-stage GMM classifier is applied. The final correct gender identification rate is 96.4%. There are 108 out of 111 male speakers and 106 out of 111 female speakers whose gender are correctly identified.

From the experimental results of the AWGN case, we are able to say that our proposed system is robust to AWGN. Furthermore, we can know that the performance of our proposed system in the 5dB case is better than the one in the 0dB case. When the noise becomes stronger, the performance will definitely degrades. Compare the two experiments in both 5dB and 0dB case, we may find out that no matter we use clean training samples or corrupted training samples for training, the excellent performance remains. It indicates a big advantage of the potential application.

From all the experimental results of the noise scenario shown above, we are able to say that our proposed system exhibits the characteristics of being robust to the background noise and strong AWGN. The reason why our proposed system is robust to the noise is that only the envelope information of the speech signal is used for pitch period estimation and a bandpass filtering process involves in the estimation method. For envelope signal, the noise effect is less severe than the original speech signal. The energy of the noise usually spreads at the whole spectrum band or concentrates within a certain frequency band which is different from the frequency band of the speech signal. Thus, the bandpass filtering with the formant frequency as the center frequency will remove a significant part of noise energy but preserve the most energy of the formant resonances. The noise effect can be greatly reduced and the pitch feature can be accurately extracted.

VI. Conclusions

For speech-based gender identification, the most commonly used features are pitch period and MFCC. MFCC feature usually requires higher computation complexity, providing unnecessary linguistic information for gender identification. Also MFCC feature is not robust to the change of the recording condition. In this paper, a new pitch-based gender identification system with suspicious speaker detection scheme is proposed. In our system, the short-time speech is modeled as sums of AM-FM formant models and

the pitch feature is extracted by estimating the periods of the AM components. Since only the envelope information of the original speech signal is used and a bandpass filtering process involves in the pitch feature extraction, the noise effect and the distortion effect caused by the change of the recording condition would be greatly reduced. Thus, the pitch feature used in the proposed system is believed to be language-independent, microphone-independent and robust to the noise condition. On the other hand, to address the mis-identification problem caused by the overlap of the pitch values between male and female voices, in our system, a two-stage classifier with suspicious speaker detection scheme is adopted. The first stage classifier adopts thresholding method to do the gender identification for those speakers whose pitch values do not fall in the overlap. All the rest speakers whose pitch values fall in the overlap are declared as the suspicious speakers and their gender decision is made by the second-stage classifier using GMM method. In this way, the mis-identification problem caused by the overlap of pitch values could be solved and the computation complexity is reduced at the utmost. The excellent performance of the proposed gender identification system is validated by the experimental results. The experimental results also indicate that the proposed gender identification system exhibits the characteristics of language/content independence, microphone independence and being robust to the background noise and strong AWGN.

ACKNOWLEDGMENTS

The authors would like to thank Ram Keralapura for suggestions that helped to improve the presentation of this paper.

REFERENCES

[1] H. Hanson and E. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *The Journal of the Acoustical Society of America*, vol. 106, p. 1064, 1999.

[2] A. Kondoz, *Digital speech: coding for low bit rate communication systems*. John Wiley and Sons Ltd, 2004.

[3] A. Acero and X. Huang, "Speaker and gender normalization for continuous-density hidden Markov models," in *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*, vol. 1. Citeseer, 1996.

[4] C. Neti and S. Roukos, "Phone-context specific gender-dependent acoustic-models forcontinuous speech recognition," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding, 1997. Proceedings.*, 1997, pp. 192–198.

[5] E. Parris and M. Carey, "Language independent gender identification," in *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*, vol. 2, 1996.

[6] D. Deepawale, R. Bachu, and B. Barkana, "Energy Estimation between Adjacent Formant Frequencies to Identify Speaker's Gender," in *Proceedings of the Fifth International Conference on Information Technology: New Generations*. IEEE Computer Society, 2008, pp. 772–776.

[7] H. Harb and L. Chen, "Voice-based gender identification in multimedia applications," *Journal of Intelligent Information Systems*, vol. 24, no. 2, pp. 179–198, 2005.

[8] M. Gelfer and V. Mikos, "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, 2005.

[9] W. Hess, *Pitch determination of speech signals: algorithms and devices*. Springer, 1983.

[10] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 727–730, 2001.

[11] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE transactions on acoustics, speech and signal processing*, vol. 22, no. 5, pp. 353–362, 1974.

[12] E. Yucesoy and V. Nabiyev, "Gender identification of the speaker using DTW method," in *Proceedings of the 2009 IEEE 17th Signal Processing and Communications Applications Conference*, 2009, pp. 273–276.

[13] M. Reflection and L. Coefficients, "AUTOMATIC GENDER IDENTIFICATION UNDER ADVERSE CONDITIONS."

[14] H. Ting, Y. Yingchun, and W. Zhaohui, "Combining MFCC and pitch to enhance the performance of the gender recognition," in *Signal Processing, 2006 8th International Conference on*, vol. 1, 2006.
[15] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on signal processing*, vol. 41, no. 10, pp. 3024–3051, 1993.
[16] P. Tsiakoulis and A. Potamianos, "Statistical analysis of amplitude modulation in speech signals using an AM-FM model," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing-Volume 00*. IEEE Computer Society, 2009, pp. 3981–3984.
[17] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Transactions on acoustics, speech and signal processing*, vol. 22, no. 2, pp. 135–141, 1974.
[18] C. Espy-Wilson, "A phonetically based semivowel recognition system," in *Proceedings of the 1986 International Conference on Acoustic Speech and Signal Processing*, 1986, pp. 2775–2778.
[19] J. Tierney, "A study of LPC analysis of speech in additive noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 389–397, 1980.
[20] S. Furui, *Digital speech processing, synthesis, and recognition*. Marcel Dekker Inc, 1989.
[21] A. Watanabe, "Formant estimation method using inverse-filter control." Institute of Electrical and Electronics Engineers.
[22] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 313–327, 1998.
[23] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 36–48, 1998.
[24] R. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 129–134, 1993.
[25] P. Maragos, J. Kaiser, and T. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on signal processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

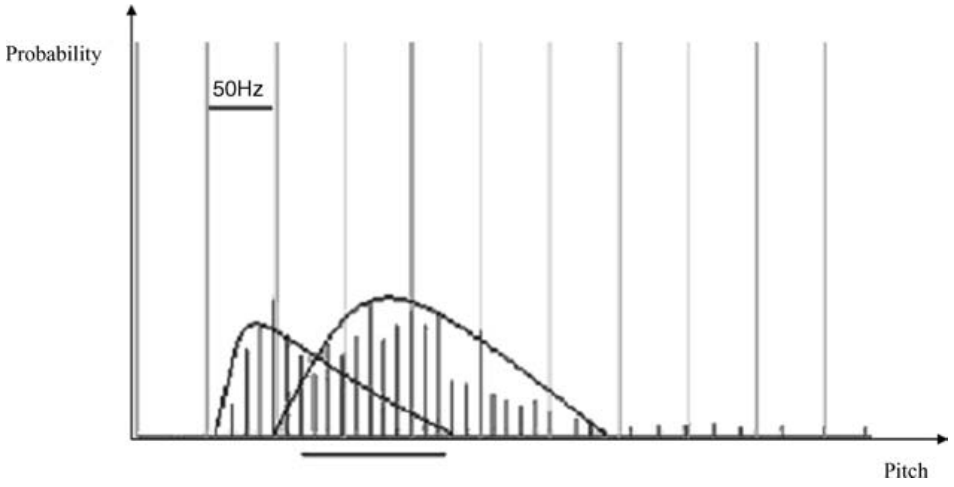

Fig. 1. Overlap of Pitch Value Histogram for Male Speakers and Females Speakers

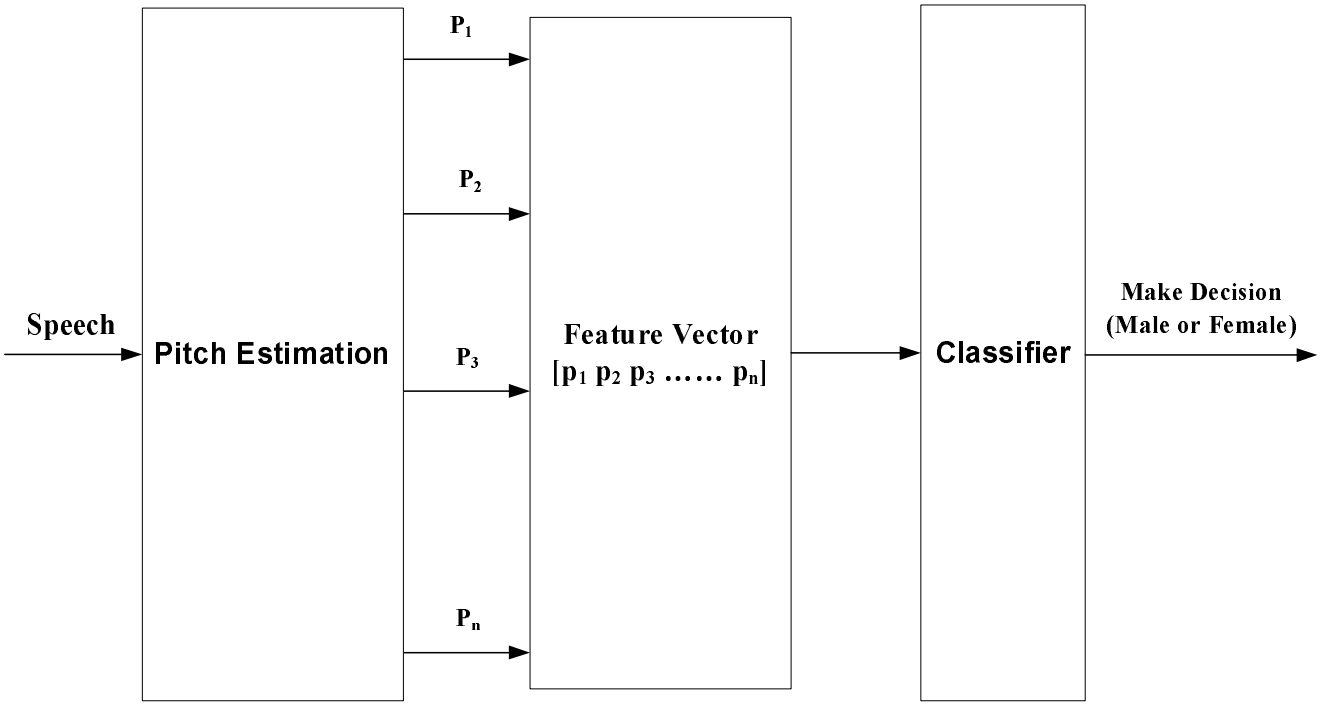

Fig. 2. Overall Architecture of Gender Identification System

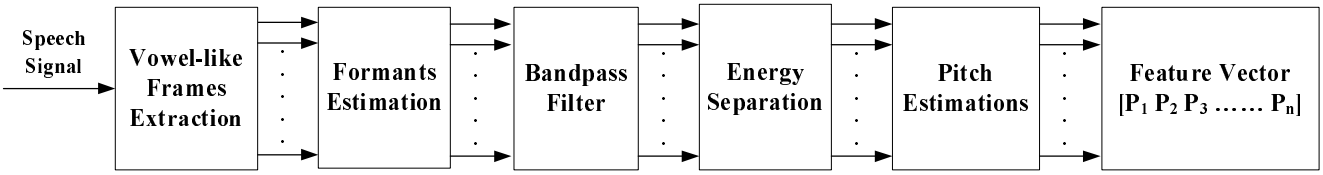

Fig. 3. Structure of Pitch Period Estimator

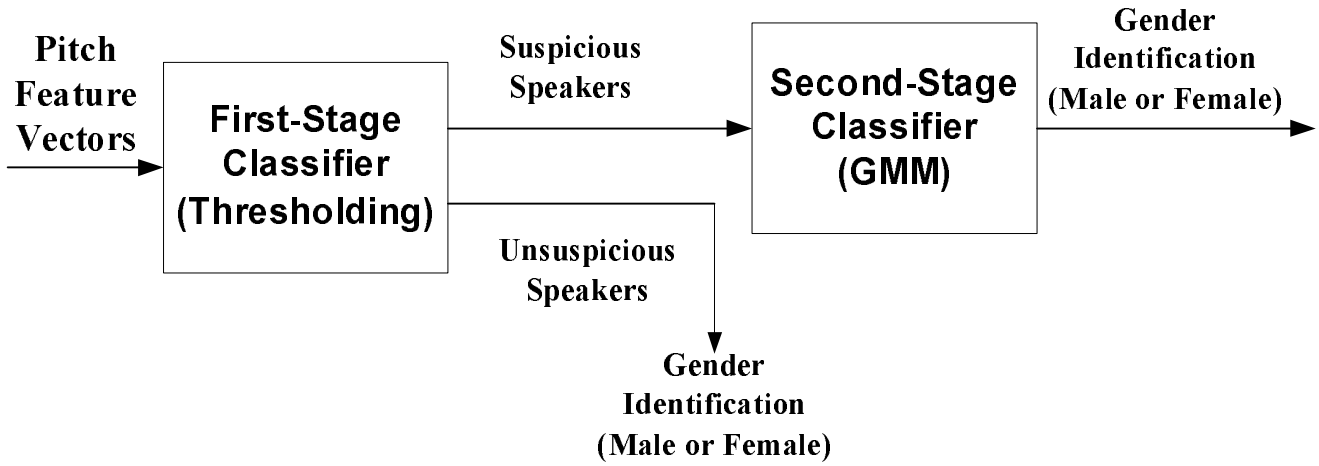

Fig. 4. Structure of Two-Stage Classifier

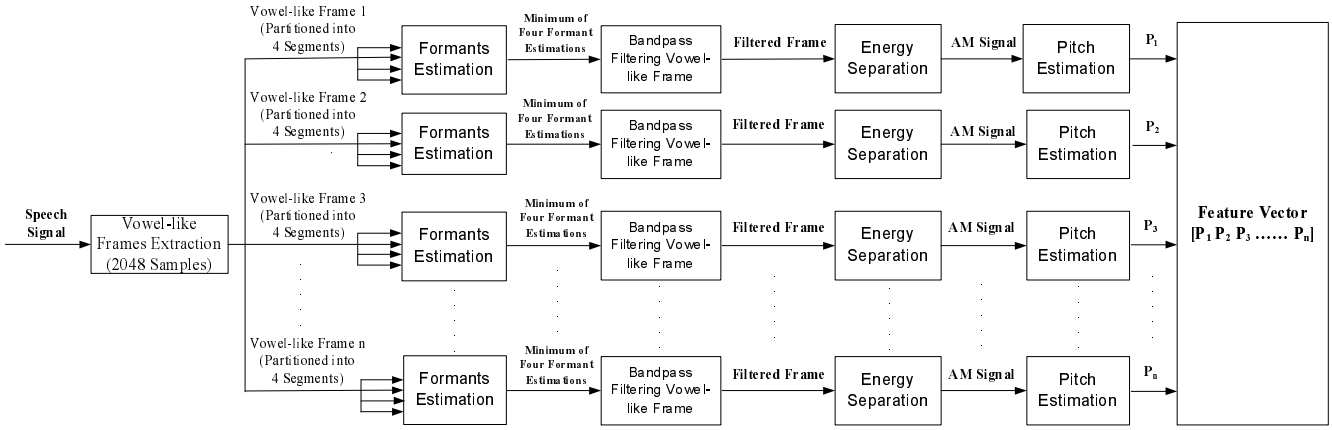

Fig. 5. Flow Chart of Pitch Feature Extraction

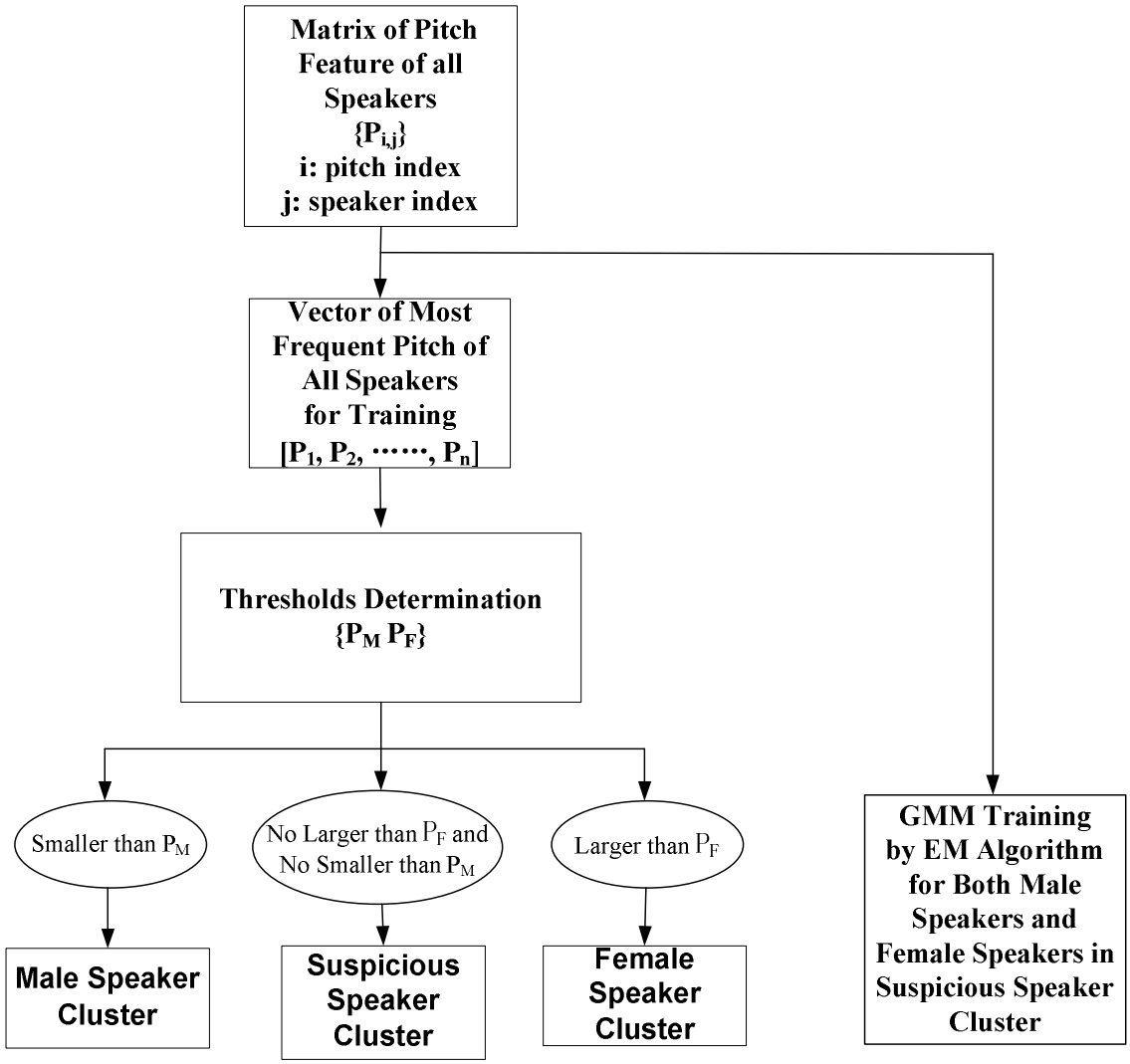

Fig. 6. Two-Stage Classifier in Training Phase

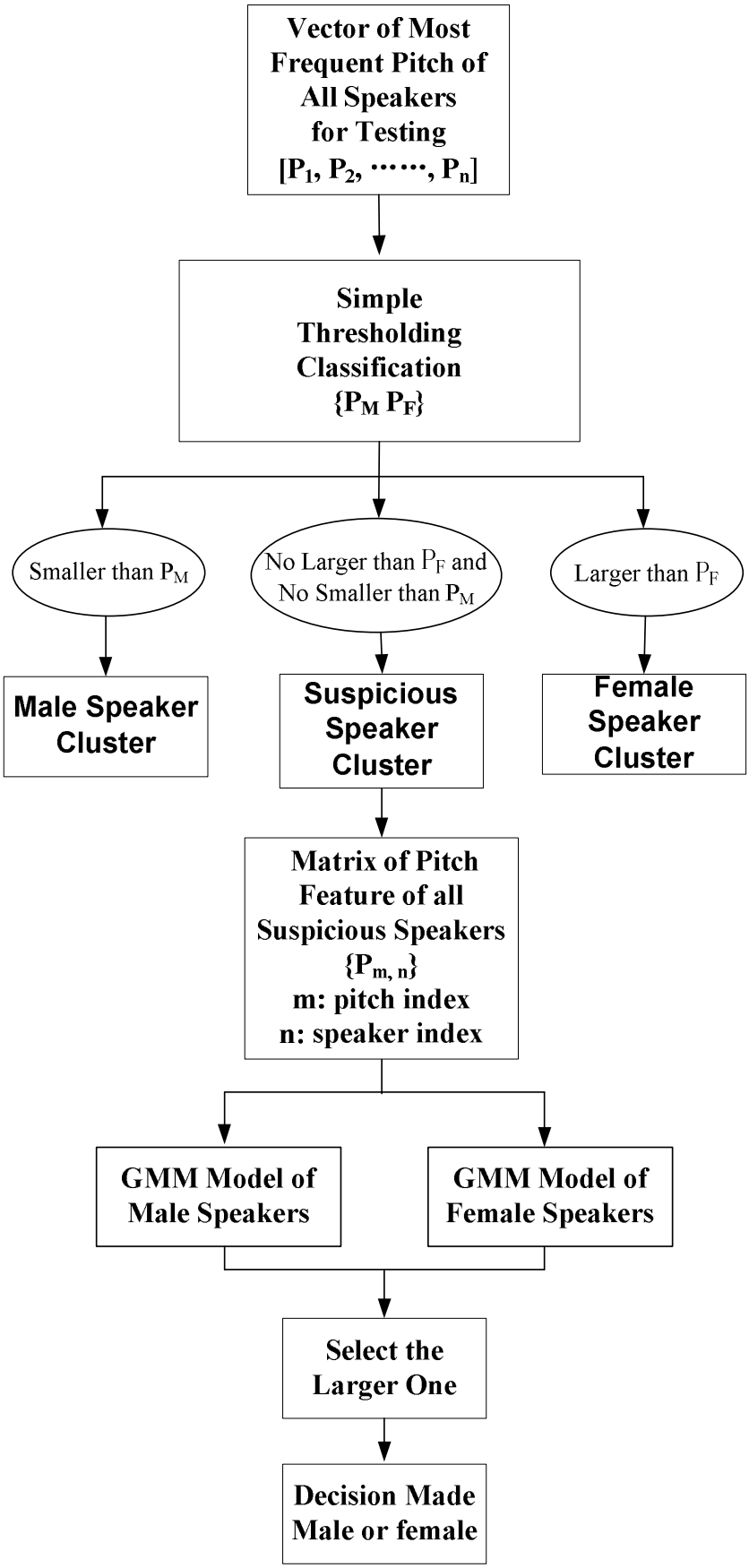

Fig. 7. Two-Stage Classifier in Testing Phase