Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006

# ROBUST GMM BASED GENDER CLASSIFICATION USING PITCH AND RASTA-PLP PARAMETERS OF SPEECH

**YU-MIN ZENG[1, 2], ZHEN-YANG WU[1], TIAGO FALK[3], WAI-YIP CHAN[3]**

[1]Department of Radio Engineering, Southeast University, Nanjing 210096, China
[2]School of Physics Science and Technology, Nanjing Normal University, Nanjing 210097, China
[3]Department of Electrical and Computer Engineering, Queen's University, Kingston, K7L 3N6, Canada
E-MAIL: zengyumin@njnu.edu.cn, zhenyang@seu.edu.cn

**Abstract:**

**A novel gender classification system has been proposed based on Gaussian Mixture Models, which apply the combined parameters of pitch and $10^{th}$ order relative spectral perceptual linear predictive coefficients to model the characteristics of male and female speech. The performances of gender classification system have been evaluated on the conditions of clean speech, noisy speech and multi-language. The simulations show that the performance of the proposed gender classifier is excellent; it is very robust for noise and completely independent of languages; the classification accuracy is as high as above 98% for all clean speech and remains 95% for most noisy speech, even the SNR of speech is degraded to 0dB.**

**Keywords:**

**Gender classification; GMM; RASTA-PLP**

## 1. Introduction

Much information can be inferred form a speech, such as sequences of words, gender, age, dialect, emotion, ethnicity, and even level of education, height or weight etc. Gender is an important characteristic of a speech. Automatic gender classification is a technique that aims to determine the sex of the speaker through speech signal analysis. Automatically detecting the gender of a speaker has several potential applications such as (1) sorting telephone calls by gender (e.g. for gender sensitive surveys), (2) as part of an automatic speech recognition system to enhance speaker adaptation, and (3) as part of automatic speaker recognition systems. In the past, many methods of gender classification have been proposed. For parameters selections, some methods used gender dependent features, such as pitch and formant [1][2], some applied general speech features used by general pattern recognition, such as autocorrelation coefficients, log area ratios, reflection coefficients, LPC, MFCC etc. [2][3], and some combined pitch and general speech features [4-6]. The Artificial Neural Network (ANN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Bayes and Support Vector Machine (SVM) were used for the classifier respectively. Most of the gender classifications were tested for clean speech and the accuracy remains below 95%.

In this paper, we propose a gender classification system based on GMM, which apply the combined parameters of pitch and relative spectral perceptual linear predictive (RASTA-PLP) coefficients to model the behavior of male and female speech. We train several GMMs with different components and covariance matrices, and test the performances of the gender classifier under the conditions of clean speech and various degraded speech. We also experiment with speech files in different languages. The simulation results show that the gender classifier we proposed provides very good performances both for clean speech and noisy speech, and is independent of languages.

## 2. Gender classification method

### 2.1. Features selection and extraction

It is well known that the major difference between male and female speech is the pitch. Generally, women have higher pitch than men, as shown in Figure. 1. This figure shows average pitch values of 2052 different speech files for both male and female speakers. The vertical axis represents pitch values in Hz and the horizontal axis represents the speech file number. This discrimination qualifies pitch as an effective feature for gender classification. The accurate pitch extraction is not an easy task due to the non-stationarity and quasi-periodicity of speech signal, as well as the interaction between the glottal excitation and the vocal tract. The autocorrelation pitch detector was shown to be more robust to noise. In our classification system, the modified autocorrelation method [7] is used to estimate the pitch of segmented speech. And then, the estimated results are further filtered by a medium

filter in order to increasing the robustness of pitch

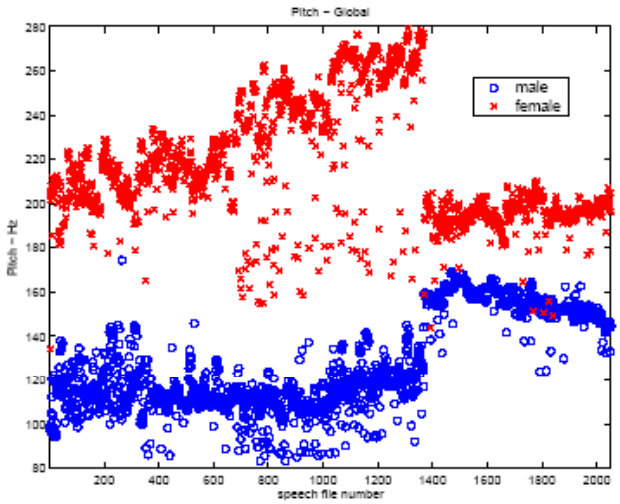


Figure 1. Pitch of 2052 speech files for male and female speakers

estimation.

However, there is some overlap for the pitches of high-pitch males and low-pitch females, as can be seen in Figure. 1. For increasing classification accuracy, combining pitch and other speech feature is necessary. The perceptual linear predictive (PLP) uses the concepts from the psychoacoustics of hearing and yields lower dimension representation of speech, which is found to be useful for automatic speech recognition. Along with PLP, relative spectral (RASTA) method uses filtering in the log domain of the power spectrum to compensate for the channel effects in recognizers, which is proved to be more robust for noisy speech recognition. So that, the relative spectral perceptual linear predictive (RASTA-PLP) coefficients [8] combined with pitch are chosen for the parameters of our proposed gender classification system. The extraction of

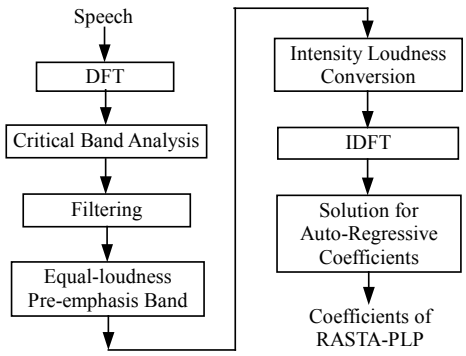


Figure. 2 Process of calculation of

RASTA-PLP is illustrated in Figure 2.

## 2.2. Gaussian Mixture Model and EM algorithm

Gaussian mixture models (GMMs) have been used extensively in speech processing. The reasons are: (1) univariate Gaussian densities have a simple and concise representation, depending uniquely on two parameters, mean and variance, and (2) the Gaussian mixture distribution is universally studied and its behaviors are widely known.

In principle, GMM can approximate any probability density function to an arbitrary accuracy. Let $\mathbf{u}$ be a $K$-dimensional vector, a Gaussian mixture density is a weighted sum of $M$ component densities

$$p(\mathbf{u} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = \sum_{i=1}^{M} \alpha_i \cdot b_i(\mathbf{u}) \tag{1}$$

where $\alpha_i \geq 0, i = 1, \cdots, M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{u}), i = 1, \cdots, M$ are the $K$-variate Gaussian densities

$$b_i(\mathbf{u}) = \frac{1}{(2\pi)^{K/2} \mid \boldsymbol{\Sigma}_i \mid^{1/2}} \exp\left( -\frac{1}{2} (\mathbf{u} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{u} - \boldsymbol{\mu}_i) \right) \tag{2}$$

with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$.

GMMs can assume several different forms, depending on the type of covariance matrices. The two most widely used are full and diagonal covariance matrices.

In our experiment, the training of GMMs is an unsupervised learning. The expectation-maximization (EM) algorithm is used to train the GM densities [9]. The EM algorithm iterations produce a sequence of GM models with monotonically non-decreasing (log-) likelihood values. Giving the training vector U={$\mathbf{u}_1$, $\mathbf{u}_2$, …, $\mathbf{u}_N$}, the EM ($k$+1)th iteration computations are

$$h_{ij} = \frac{\mid \hat{\boldsymbol{\Sigma}}_j^{(k)} \mid^{-1/2} \exp\left( -\frac{1}{2} (\mathbf{u}_i - \hat{\boldsymbol{\mu}}_j^{(k)})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_j^{(k)^{-1}} (\mathbf{u}_i - \hat{\boldsymbol{\mu}}_j^{(k)}) \right)}{\sum_{l=1}^{M} \left( \mid \hat{\boldsymbol{\Sigma}}_l^{(k)} \mid^{-1/2} \exp\left( -\frac{1}{2} (\mathbf{u}_i - \hat{\boldsymbol{\mu}}_l^{(k)})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_l^{(k)^{-1}} (\mathbf{u}_i - \hat{\boldsymbol{\mu}}_l^{(k)}) \right) \right)} \tag{3}$$

$$\hat{\alpha}_j^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} h_{ij}, \; j = 1,2,\cdots,M \tag{4}$$

$$\hat{\boldsymbol{\mu}}_j^{(k+1)} = \sum_{i=1}^{N} h_{ij} \mathbf{u}_i \Big/ \sum_{i=1}^{N} h_{ij}, \; j = 1,2,\cdots,M \tag{5}$$

$$\hat{\boldsymbol{\Sigma}}_j^{(k+1)} = \frac{\sum_{i=1}^{N} h_{ij} (\mathbf{u}_i - \hat{\boldsymbol{\mu}}_j^{(k+1)})(\mathbf{u}_i - \hat{\boldsymbol{\mu}}_j^{(k+1)})^{\mathrm{T}}}{\sum_{i=1}^{N} h_{ij}}, \; j = 1,2,\cdots,M \tag{6}$$

Though the EM algorithm converges to a maximum likelihood, depending on the algorithm initialization values,

it may converge to a local maximum and not the global maximum. Here we use the *k-means* algorithm to initialize the GMM parameters.

Gaussian mixture densities are used to model the 12-dimensional feature vectors comprised of the pitch and the 10th order RASTA-PLP coefficients. We only consider the features that are extracted from voiced speech frames. Using clean and degraded speech signals, we train two different Gaussian mixture densities for male and female speech, $p_{male}(\mathbf{u}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha})$ and $p_{female}(\mathbf{u}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha})$.

### 2.3. Gender classification

In practice, for a given speech signal, a sequence of the feature coefficients vector can be obtained. The length of feature sequence is dependent on the length of the speech file and the number of voiced segments in the file. The log-likelihood is commonly used to measure how well a GM model fits to experimental test data. Assuming independence of the vectors between frames and the feature sequence being $\mathbf{x}=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, the likelihood probability can be expressed as

$$p_{gender}(\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha}) = \prod_{i=1}^{N} p_{gender}(\mathbf{x}_i\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha}) \qquad (7)$$

where the subscript "*gender*" represents either "male", or "female". The normalized log-likelihood is expressed as

$$LL_{gender}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\log(p_{gender}(\mathbf{x}_i\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha})) \qquad (8)$$

Given an utterance from an unknown speaker, the normalized log-likelihood of $LL_{male}(\mathbf{x})$ and $LL_{female}(\mathbf{x})$ are calculated. If $LL_{male}(\mathbf{x}) > LL_{female}(\mathbf{x})$, the speaker is determined as a "male", otherwise, the speaker is determined as a "female".

## 3. Experiments

Several experiments were carried out to evaluate the proposed gender classifier for various classification conditions. A total of four databases, such as "TIMIT" etc, comprised of four multilingual (English, German, Japanese, Italian) speech signals and their degraded speech signals, were used for training and testing to our classifier. The training databases were 182 male speakers (132 English, 30 German, 10 Japanese and 10 Italian), totally 2510 speech files (1670 clean files and 840 noisy files), and 133 female speakers (93 English, 20 German, 10 Japanese and 10 Italian), totally 1850 speech files (1230 clean and 620 noisy). The testing databases were 94 male speakers (64 English, 20 German, 5 Japanese and 5 Italian), totally 6720 files (840 clean and 5880 noisy), and 51 female speakers (31 English, 10 German, 5 Japanese and 5 Italian), totally 5060 files (460 clean and 4600 noisy). The noisy speeches were obtained from part of clean speeches by adding in some white, babble, factory and car noises respectively. All speech signals were 8kHz sample rate and 16bits/sample. The pitch and RASTA-PLP coefficients were extracted every 10ms segment (only the parameters of voiced speech were kept). The 10th order of RASTA-PLP was applied. So that the dimension of parameter vector and GMM is $K$=12 (1 Pitch + 11 RASTA-PLP).

The GMMs of both full and diagonal covariance matrices with several different numbers of components were trained respectively. We tested all trained-types of GM model in clean speech condition. Table 1 shows the classification accuracies for clean speech with mixed languages. In the table, "Diag_4" denotes the type of GMM is diagonal covariance matrix with 4 components, and "Full_2" denotes the type of GMM is full covariance matrix with 2 components. The results show the performance of the proposed gender classifier is excellent for clean speech; its classification accuracy is above 98%. The results also indicate that large number of components of GMM is not necessary; 4 to 8 components are good enough for our proposed method.

Table 1. The classification accuracies with different GMMs for clean speeches

| | Diag_4 | Diag_6 | Diag_8 | Diag_12 | Diag_16 |
|---|---|---|---|---|---|
| Male | 97.7% | 97.7% | 98.0% | 98.0% | 97.9% |
| Female | 98.6% | 98.5% | 98.5% | 98.6% | 98.7% |
| Total | 98.1% | 98.0% | 98.1% | 98.2% | 98.2% |
| | Full_2 | Full_4 | Full_6 | Full_8 | Full_10 |
| Male | 97.8% | 97.9% | 98.0% | 97.9% | 98.0% |
| Female | 98.3% | 98.4% | 98.7% | 98.6% | 98.7% |
| Total | 97.9% | 98.0% | 98.2% | 98.1% | 98.2% |

Table 2. The total accuracies of classification for different noisy speeches

| | SNR | white | babble | factory | car | total |
|---|---|---|---|---|---|---|
| Diag_8 | ∞ | | | | | 98.1% |
| | 20dB | 98.1% | 97.4% | 98.0% | 97.9% | 97.9% |
| | 10dB | 97.5% | 96.9% | 97.7% | 97.3% | 97.4% |
| | 0dB | 95.6% | 93.9% | 95.0% | 95.3% | 95.0% |
| Full_4 | ∞ | | | | | 98.0% |
| | 20dB | 98.0% | 97.3% | 97.7% | 97.8% | 97.7% |
| | 10dB | 97.4% | 96.9% | 97.2% | 97.1% | 97.2% |
| | 0dB | 95.5% | 94.1% | 95.4% | 95.2% | 95.1% |

We have also carried out the experiment to evaluate the performances of our proposed method for various noisy

speeches and multilingual speeches. Here only two types of GMM, "Diag_8" and "Full_4", were used for testing. Table 2 presents the total testing results for several kinds of noisy speech with different Signal-to-Noise Ratio. And Table 3 shows the accuracies of classification for different language using the GMM of "Diag_8". From the tables it can also be seen that our proposed classifier is a very robust gender classifier, the accuracy remains above 95% for most kinds of noisy speech, even the SNR is decreased to 0dB and the classifier is completely independent of languages.

Table 3. The accuracies of classification for different language noisy speeches

| Languages | | English | German | Japanese | Italian | total |
|---|---|---|---|---|---|---|
| SNR | ∞ | 98.1% | 98.0% | 98.2% | 98.3% | 98.1% |
| | 20dB | 97.9% | 98.0% | 98.0% | 97.8% | 97.9% |
| | 10dB | 97.5% | 97.3% | 97.3% | 97.5% | 97.4% |
| | 0dB | 94.9% | 95.1% | 95.2% | 95.3% | 95.0% |

Another experiment was also employed. The 8-component GMM with diagonal covariance matrix, "Diag_8", only using Pitch or RASTA-PLP for modeling, was trained respectively. The performances of this GMM were evaluated in the conditions of clean and noisy speeches with multi-language, just for comparing to the proposed method. The total accuracies of classification are presented in Table 4. It is evident that the classification performance by using combining parameters of Pitch and RASTA-PLP is obviously better than that by only using the parameter of Pitch or RASTA-PLP.

Table 4. The comparison of classification accuracy for three types of GMM

| | ∞ | 20dB | 10dB | 0dB |
|---|---|---|---|---|
| P | 96.5% | 96.1% | 95.4% | 91.5% |
| R-PLP | 93.7% | 92.6% | 89.2% | 76.1% |
| P & R-PLP | 98.1% | 97.9% | 97.4% | 95.0% |

## 4. Conclusions

In this paper, a novel gender classifier has been proposed based on Gaussian Mixture Models. The combination of pitch and $10^{th}$ order RASTA-PLP coefficients have been used to model the characteristics of male and female speech by means of two GMMs of male and female. Several GMMs with different covariance matrices and components have been trained, and the classification system has been tested on the conditions of clean speech, noisy speech and multi-language. The experimental results show that 4 - 8 components of GMM is sufficient for our proposed method; the performance of the proposed gender classifier is excellent, it is very robust for noise and is independent of language; the classification accuracy is above 98% for all clean speech and remains 95% for most kinds of 0dB noisy speech.

**Acknowledgements**

This work is supported by the 973 Program of China under Grant No. 2002CB312102 and the College Natural Science Research Program of Jiangsu Province of China.

**References**

[1] R. Vergin, A. Farhat and D. O'Shaughnessy, "Robust gender-dependent acoustic-phonetic modeling in continuous speech recognition based on a new automatic male/female classification", Proc. of IEEE Int. Conf. on Spoken Language (ICSLP), pp. 1081-1084, Oct. 1996.

[2] S. Slomka and S. Sridharan, "Automatic gender identification optimized for language independence", Proc. of IEEE TENCON'97, pp. 145-148, Dec. 1997.

[3] K. Wu and D. Childers, "Gdener recognition from speech. Part I: Coarse analysis", J. Acoust. Soc. Am., Vol. 90, No. 4, pp. 1828-1840, 1991.

[4] E. Parris and M. Carey, "Language Independent gender identification", Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 685-688, May 1996.

[5] H. Harb and L. Chen, "Gender identification using a general audio classifier", Proc. of Int. Conf. on Multimedia and Exposition (ICME), Vol. 2, pp. 733-736, July 2003.

[6] I. Shafran, M. Riley, and M. Mohri, "Voice signatures", Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 31-36, Dec. 2003.

[7] L. Rabiner, et al., "A Comparative Study of Several Pitch Detection Algorithms", IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 5, pp. 399-417, Oct. 1976.

[8] H. Hermansky and N. Morgan, "RASTA Processing of Speech", IEEE Trans. Speech and Audio Processing, Vol. 2, No. 4, pp. 578-589, Oct. 1994.

[9] A. Dempster, N. Lair and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. Royal Statistical Society, Vol. 39, pp. 1-38, 1977.