Method for Speaker Gender Classification Based on Gaussian Mixture Modeling

Vasyl Semenov
EPAM Department of Information Technologies
American University Kyiv,
Kyiv Academic University
Kyiv, Ukraine
vasyl.delta@gmail.com

Yevheniya V. Semenova
Laboratory of Data Science and Machine Learning
Kyiv Academic University,
Institute of Mathematics of NANU
Kyiv, Ukraine
semenovaevgen@gmail.com

***Abstract*— The automatic gender identification is an important problem both as independent task and as a component of different natural language processing (NLP) systems. In this paper the method for automatic speaker gender classification is proposed and its basic algorithmic stages are described. The method is based on the modeling of voice acoustic parameters distribution by weighted sum of several Gaussian distributions (Gaussian Mixture Modeling, GMM). The set of cepstral RASTA-PLP coefficients extended by fundamental frequency was selected as the vector of acoustic features. GMMs for male and female speakers were trained by Expectation-Maximization (EM) method with initialization by K-means algorithm. The dependency of classification accuracy on the GMM types (with diagonal and full-size covariance matrices) as well as their orders was investigated. In different experiments proposed method has shown classification accuracy from 91% to 100%. The comparison of proposed method both with logistic regression and five-layer neural network is also given.**

Keywords—cepstral coefficients, pitch frequency, Gaussian mixture models, logistic regression, neural network.

I. Introduction

The task of speaker gender identification [1, 2] is relevant for systems of automatic classification of speech information, since preliminary gender identification provides more accurate adjustment of the recognition system. In addition, speaker gender identification can be of independent interest for systems that provide law enforcement, information gathering for advertising purposes etc. Nowadays, it is an important part of NLP systems [3].

The key issues for building any recognition system are as follows:

1. Selection of features, i.e. parameters characterizing the objects to be classified (in this case, male/female voices);

2. Selection of a model, according to which the recognition system is trained and the subsequent classification is performed.

According to this principle, at the preliminary stage, feature vectors are extracted from the database of training data. Then the obtained array of features is preprocessed and used to train the classification model, resulting in some classes of features. During the testing mode, the test data are compared with the reference values obtained at the preliminary stage and thus the corresponding classification decisions are made (in the case of deep learning the classes or reference values may not be directly present).

In speaker verification systems the features are often represented by vector of cepstral parameters calculated at each frame of the speech signal [4, 5]. In this investigation we selected a set of 10 RASTA-PLP cepstral coefficients, augmented with the fundamental frequency (pitch), as a feature vector. The cepstral coefficient responsible for the signal level was excluded, i.e. the total dimensionality of the feature vector was 11.

Different approaches to classification are used in recognition tasks: Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), deep learning (DL) and others. Since we are not interested in the analysis of dynamic change of features, in this investigation we used the GMM methodology for the task of gender classification. We also compare its performance with logistic regression and five-layer neural network as presented in [2].

II. Features Calculation

An important feature used to distinguish between male and female voices is the fundamental frequency. This parameter characterizes the frequency of vocal cords oscillation during the pronunciation of sounds. To calculate the fundamental frequency, we used the autocorrelation method described in [6].

As a rule, men are characterized by lower values of the fundamental frequency compared to women. However, as can be seen from Fig. 1, these ranges do overlap, so that in some cases a female voice may correspond to a lower fundamental frequency. Therefore, the most challenging situations are when it is necessary to identify a female with a low pitch or, conversely, a male with a high pitch. In such cases, correct identification should be achieved by using parameters that reflect the differences in vocal tract structure between males and females.

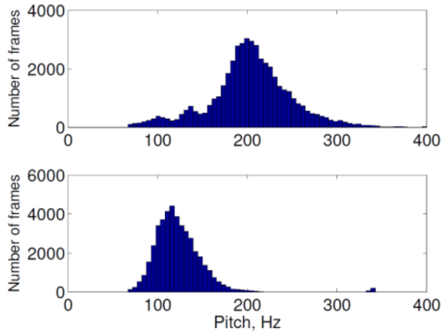

Рис. 1.Histograms of fundamental frequency (pitch) distribution for women (upper graph) and for men (lower graph).

Based on the above, we added to the feature vector 10 RASTA-PLP ("RelAtive SpecTrA") coefficients that determine the shape of the vocal tract during pronunciation of sounds. The RASTA-PLP methodology for analyzing speech signals consists of two parts - PLP (Perceptual Linear Prediction), linear prediction taking into account the features of auditory perception and RASTA, processing designed to remove spectral components, which are not typical for the dynamics of speech signal [4, 5].

III. Gaussian mixture model

The basic idea of the GMM modeling is to represent the distribution density of the feature vector (of dimension d) as a weighted sum of Gaussian distribution densities [7]:

$$p(\mathbf{x}) = \sum_{i=1}^{M} \alpha_i b(\mathbf{x}/\boldsymbol{\mu}_i, \mathbf{D}_i)\ , \qquad (1)$$

where $\alpha_i, i=1,...,M$ are weighting coefficients and $b(\mathbf{x}/\boldsymbol{\mu},\mathbf{D})$ is a Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{D}$:

$$b(\mathbf{x}/\boldsymbol{\mu},\mathbf{D}) = \frac{1}{\sqrt{2\pi \det \mathbf{D}}} \times \exp[-0.5(\mathbf{x}-\boldsymbol{\mu})^T\, \mathbf{D}^{-1}(\mathbf{x}-\boldsymbol{\mu})].$$

In fact, representing the density as a sum of Gaussians corresponds to partitioning the set of acoustic parameters into M subclasses [8].

GMMs must be independently trained for each of the alternative classes. This means that a different set of parameters $\boldsymbol{\lambda} = \{\alpha_i,\ \boldsymbol{\mu}_i,\ \mathbf{D}_i,\ i=1,...,M\}$ must be found for each class. The input for training is a set of acoustic feature vectors $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$.

The determination of these parameters requires maximizing the maximum likelihood functional

$$p(\mathbf{X}/\boldsymbol{\lambda}) = \prod_{t=1}^{T} p(\mathbf{x}_t/\boldsymbol{\lambda}). \qquad (2)$$

Maximizing the function (2) is not possible analytically. Therefore, EM (Expectation-maximization) algorithm is used to iteratively maximize it [9, 10].

Below are the equations for the iterative calculation of the parameters $\boldsymbol{\lambda} = \{\alpha_i, \boldsymbol{\mu}_i, \mathbf{D}_i, i=1,\ldots,M\}$ in the case of diagonal covariance matrices [10] (i.e. when $\mathbf{D}_i = \mathrm{diag}\{\sigma_i^1, \sigma_i^2, ..., \sigma_i^d\}$):

- update of a posteriori probabilities of belonging to the *i*-th class:

$$P(\mathbf{x}_t \in \mathbf{C}_i) = \frac{\alpha_i b_i(\mathbf{x}_t)}{\sum_{j=1}^{M} \alpha_j b_j(\mathbf{x}_t)},$$

where

$$b_i(\mathbf{x}_t) = \frac{\exp\left\{-\frac{1}{2}\sum_{k=1}^{d}\frac{(\mathbf{x}_t^k - \boldsymbol{\mu}_i^k)^2}{(\sigma_i^k)^2}\right\}}{\prod_{k=1}^{d}\sigma_i^k};$$

- update of weights:

$$\alpha_i = \frac{1}{N}\sum_{t=1}^{T} P(\mathbf{x}_t \in \mathbf{C}_m);$$

- update of mean values:

$$\boldsymbol{\mu}_i = \frac{\sum_{t=1}^{T} P(\mathbf{x}_t \in \mathbf{C}_i)\mathbf{x}_t}{\sum_{i=1}^{N} P(\mathbf{x}_t \in \mathbf{C}_i)};$$

- update of variances:

$$(\sigma_i^k)^2 = \frac{\sum_{t=1}^{T} P(\mathbf{x}_t \in \mathbf{C}_i)(\mathbf{x}_t^k)^2}{\sum_{t=1}^{T} P(\mathbf{x}_t \in \mathbf{C}_i)} - (\boldsymbol{\mu}_i^k)^2.$$

The equations for the estimation of parameters for GMM with full-size covariance matrices are similar and can be found, e.g. in [10].

As usual, the problem of initial initialization is acute for the EM method [10]. That is why in this investigation we used the K-means algorithm [8] with uneven dichotomy to select an initial approximation for GMM parameters in the following way.

Applying the K-means algorithm to a set of acoustic feature vectors $\mathbf{X}=[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$ allows us to find M quantums that serve as initialization for the mathematical expectations $\boldsymbol{\mu}_i$, $i=1,\ldots,M$. Then, by selecting the vectors falling into the cell $\mathbf{C}_i$, we obtain an approximation for the variances

$$(\sigma_i^k)^2 = \frac{\sum_{i \in \mathbf{C}_i} (\mathbf{x}_i^k - \boldsymbol{\mu}_i^k)^2}{N^{(i)}}, \quad k = 1, \ldots, d,$$

where $N^{(i)}$ is the number of elements in the *i*-th cell.

The values α are initialized as

$$\alpha_i = \frac{N^{(i)}}{N}.$$

Correspondingly, in the test mode, to classify a set of observations $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$, hypothesis testing reduces to a comparison of the probability densities corresponding to the GMM parameters for the speakers of each gender: $p(\mathbf{Y}/\alpha^{(\mathrm{mal.})}, \boldsymbol{\mu}^{(\mathrm{mal.})}, \mathbf{D}^{(\mathrm{mal.})})$ and $p(\mathbf{Y}/\alpha^{(\mathrm{fem.})}, \boldsymbol{\mu}^{(\mathrm{fem.})}, \mathbf{D}^{(\mathrm{fem.})})$. Assuming the independence of the observation vectors, we write these values in logarithmic scale:

$$L^{(\mathrm{mal.})} = \frac{1}{N}\sum_{n=1}^{N} \log p\left(\mathbf{y}_n / \alpha^{(\mathrm{mal.})}, \boldsymbol{\mu}^{(\mathrm{mal.})}, \mathbf{D}^{(\mathrm{mal.})}\right),$$

$$L^{(\mathrm{fem.})} = \frac{1}{N}\sum_{n=1}^{N} \log p\left(\mathbf{y}_n / \alpha^{(\mathrm{fem.})}, \boldsymbol{\mu}^{(\mathrm{fem.})}, \mathbf{D}^{(\mathrm{fem.})}\right),$$

where both probability densities are calculated in accordance with expression (1).

If $L^{(\mathrm{mal.})} > L^{(\mathrm{fem.})}$, the decision is made that the current speech frame belongs to male speaker. Otherwise, it is assumed that the current frame contains female voice.

IV. Experimental results

The modeling of proposed method was done with the help of two independent based of speech signals.

- *Base 1.* 16 men and 11 women participated in the compilation of the records. The languages included Ukraine and English (USA). For each speaker 10 files were selected, with a total duration of approximately 8 minutes for men and 6 minutes for women.

- *Base 2*. This base was taken from CLSU records where 21 men and 13 women took part in the recording. The languages included were Portuguese (Brazil), English, German, Hindi, Hungarian, Japanese, Spanish and Russian. The total duration was 20 minutes for men and women (103 and 154 files, respectively).

In the first experiment, base 1 was used as a training set and base 2 as a test set. In the second experiment, respectively, base 2 was used as the training set and base 1 as the test set. The number of GMM components was set to 1, 4, 8 and 16.

Tables 1 and 2 show the classification accuracy for different orders of Gaussian mixtures and types of covariance matrices for the first and second experiments, respectively. The lower number of errors when using base 2 as a training set can be explained by its large volume and greater diversity of speakers compared to base 1.

ТАБЛИЦЯ I. PERCENTAGE OF ERRORS FOR DIAGONAL AND FULL COVARIANCE MATRICES OF DIFFERENT DIMENSIONS (FIRST EXPERIMENT)

	Diag. 1	Diag. 4	Diag. 8	Diag. 16
Male	92.3%	96.1%	99.0%	96.1%
Female	90.9%	92.2%	92.2%	92.9%
Average	91.1%	93.8%	94.9%	94.2%
	Full 1	Full 4	Full 8	Full 16
Male	96.1%	98.1%	96.1%	97.1%
Female	95.5%	93.5%	92.9%	92.9%
Average	95.7%	95.3%	94.2%	94.6%

ТАБЛИЦЯ II. PERCENTAGE OF ERRORS FOR DIAGONAL AND FULL COVARIANCE MATRICES OF DIFFERENT DIMENSIONS (SECOND EXPERIMENT)

	Diag. 1	Diag. 4	Diag. 8	Diag. 16
Male	99.3%	100.0%	100.0%	100.0%
Female	99.1%	100.0%	100.0%	100.0%
Average	99.2%	100.0%	100.0%	100.0%
	Full 1	Full 4	Full 8	Full 16
Male	99.3%	99.3%	99.3%	99.3%
Female	100.0%	100.0%	100.0%	100.0%
Average	99.6%	99.6%	99.6%	99.6%

We also note that almost all errors in the first experiment occurred for two speakers: a female Japanese speaker with an average fundamental frequency of about 135 Hz and a male English speaker with an average fundamental frequency of about 175 Hz. These frequencies are somewhat atypical for the respective genders.

For the sake of comparison, we also calculated the accuracies obtained by logistic regression and five-layer neural network as given in [2]. For the first experiment these accuracies were 87.0 and 87.9% respectively and for the second experiment – 91.0 and 93.0% respectively. The lower accuracy obtained by neural network should not be the sign for the drawback of this approach. More precise adjustment of structure/layers/activation than that in [2] may be required.

To summarize, we can say that the order and type of GMM do not significantly affect the error rate of voice gender classification. The main factor is the diversity of speakers in the training base of speech signals.

From the practical point of view, modification with diagonal 4×4 covariance matrices is preferred, as it gives an acceptable recognition and is characterized by significantly lower computational costs compared to the use of full matrices.

V. CONCLUSIONS

1. In this paper the automatic speaker gender classifier is proposed based on modeling of the voice acoustic parameters using GMM. A vector of cepstral RASTA-PLP coefficients, augmented with fundamental frequency, was chosen as a vector of acoustic features.

2. Test results show classification accuracy rates ranging from 91% to 100% depending on the type of the training and test bases, the type of GMM covariance matrices (full/diagonal) and their orders.

3. The order and type of GMM are secondary factors for correct speaker gender classification compared to the speakers' diversity in the training database of speech signals.

4. Modification with diagonal covariance matrices of small size (e.g., 4×4) seems to be the most practical, since it gives an acceptable recognition rate and is characterized by significantly lower computational costs compared to the use of full matrices.

5. The advantage in classification accuracy over both logistic regression and neural network approach [2] was shown.

REFERENCES

[1] M. Li, K. J. Han, and Narayanan S., "Automatic speaker age and gender recognition using acoustic and prosodic level in formation," Computer Speech and Language, vol. 27, 2013, pp. 151–167.

[2] M. Buyukyilmaz and A. Cibikdiken, "Voice gender recognition using deep learning," Proceedings of the Modeling, Simulation and Optimization Technologies and Applications, 2016, pp. 409–411.

[3] Sun T. et al., "Mitigating Gender Bias in Natural Language Processing: Literature Review," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1630–1640.

[4] H. Hermansky, "Perceptual Linear Prediction (PLP) analysis of speech," J. Acoust. Soc. America, vol. 87, 1990, pp. 1738–1753.

[5] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech and Audio Processing, Vol. 2, 1994, pp. 578–589.

[6] V. Semenov, "Methods for calculating and coding the parameters of autoregressive speech model when developing the vocoder based on fixed point signal process," Journal of Automation and Information Sciences, Vol. 51, 2019, pp. 30–40.

[7] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, vol. 10, 2000, pp. 19–41.

[8] R. Xu and D. Wunsch D., "Survey of Clustering Algorithms,"IEEE Transactions on Neural Networks, Vol. 16, 2005, pp. 645-678.

[9] A. Dempster, N. Lair and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. Royal Statistical Society, vol. 39, 1977, pp. 1–38.

[10] V.K. Zadiraka and V. Semenov, "Methods for the solution of systems of nonlinear algebraic equations and functions' minimization tasks: elements of theory and applications", Naukova Dumka, Kyiv, 2023.