Discriminative weight training-based optimally weighted MFCC for gender identification

Sang-Ick Kang and Joon-Hyuk Chang^{a)}

School of Electronic Engineering

Inha University,

Incheon, 402–751, Korea

a) changjh@inha.ac.kr

Abstract: In this paper, we apply a discriminative weight training to a support vector machine (SVM) based gender identification. In our approach, the gender decision rule is derived by the SVM incorporating the optimally weighted mel-frequency cepstral coefficient (MFCC) based on a minimum classification error (MCE) method which is different from the previous works in that optimal weights are differently assigned to each MFCC which is considered more realistic. According to the experimental results, the proposed approach is found to be effective for gender identification based on the SVM.

Keywords: gender identification, MFCC, MCE

Classification: Science and engineering for electronics

References

- [1] C. Neti and S. Roukos, "Phone-context specific gender-dependent acoustic-models for continuous speech recognition," *Proc. IEEE Autom. Speech Recognition Understanding Workshop*, pp. 192–198, Dec. 1997.
- [2] D. F. Marston, "Gender adapted speech coding," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 357–360, 12–15, May 1998.
- [3] H. Harb and L. Chen, "Voice-based gender identification in multimedia applications," *J. Intell. Inf. Syst.*, vol. 24, pp. 179–198, May 2005.
- [4] C. Zheng and B. Z. Yuan, "Text-dependent speaker identification using circular hidden Markov models," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, S 13.3, pp. 580–582, March 1988.
- [5] V. Tyagi and C. Wellekens, "On desensitizing the Mel-cepstrum to spurious spectral components for robust speech recognition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 529–532, 2005.
- [6] K.-H. Lee, S.-I. Kang, D.-H. Kim, and J.-H. Chang, "A support vector machine-based gender identification using speech signal," *IEICE Trans. Commun.*, vol. E91-B, no. 10, pp. 3326–3329, Oct. 2008.
- [7] H. Harb and L. Chen, "Gender identification using a general audio classifier," *Proc. IEEE Int. Conf. Multimedia & Expo*, pp. 733–736, July 2003.
- [8] S.-I. Kang, Q.-H. Jo, and J.-H. Chang, "Discriminative weight training for a statistical model-based voice activity detection," *IEEE Signal Process.*

- Lett.*, vol. 15, pp. 170–173, Feb. 2008.
- [9] Y. Kida and T. Kawahara, “Voice activity detection based on optimially weighted combination of multiple feature,” *Proc. Interspeech*, pp. 2621–2624, Sept. 2005.
- [10] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, “The OGI multi-language telephone speech corpus,” *Proc. Int. Conf. Spoken Language Process.*, vol. 2, pp. 895–898, Oct. 1992.

1 Introduction

Gender identification refers to the classical problem of distinguishing active male speech from female speech and has many applications for a variety of speech communication systems, such as speech coding and speech recognition [1, 2]. Many approaches have been devised to improve the performance of gender identification with relevant strategies such as the hidden Markov model (HMM) and the Gaussian mixture model [3, 4]. In the prevailing gender identification systems, the cepstral features such as mel-frequency cepstral coefficient (MFCC) based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency are the most commonly used because the MFCC describes the characteristics of speech signals very compactly. Although the MFCC is known to be very efficient in characterizing the human auditory system, the MFCC values are not very robust in the actual environments, and so some researchers have proposed modifications to the basic MFCC algorithm. Especially, Tyagi and Wellekens suggested a method to desensitize the MFCC coefficients to spurious low-energy spectral perturbation and reported enhanced performance for speech recognition [5].

In this regard, it has been observed in the literature that no weights are applied to the MFCCs in all mel-filter bank indexes without taking full consideration of the relative importance of the MFCCs for gender identification [6, 7]. Accordingly, in this paper, we present a novel way for gender identification to incorporate optimally weighted MFCCs based on a minimum error classification (MCE) scheme, an approach that is well known as an effective discriminative weight training [8]. Based on a number of experiments, the proposed VAD technique is found to yield a better performance than the conventional approach of adopting equally weighted MFCCs.

2 Support vector machine-based gender identification employing MFCC

In this section, we briefly review the notion of the support vector machine incorporating the MFCCs used for gender identification [6]. At first, let $\mathbf{c}^T = [c_1, c_2, \dots, c_N]^T$ be a MFCC feature vector with dimension N where T denotes the transpose of a vector on a matrix. The MFCC approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal spectrum with the help of frequency warping, allowing for better representation of speech. Specifically, there are

many complex processes for extracting cepstra from the input speech signal, including nonlinear transforms such as the Mel-scale filter bank, log spectrum and matrix operation.

Secondly, the SVM is known to make it possible to build on optimal hyperplanes separating two binary classes. In this work, given training data consisting of N -dimensional MFCC features $\mathbf{c}_i$ and the corresponding class label z_i ($\mathbf{c}_1, z_1, \dots, \mathbf{c}_l, z_l$), $\mathbf{c} \in R^N, z \in \{+1, -1\}$, the equation for the hyperplane discriminating gender is given by

$$(\boldsymbol{\lambda} \cdot \mathbf{c}) + b = 0 \quad (1)$$

where $\boldsymbol{\lambda}$ is the weighted vector, b is the bias and $(\mathbf{u} \cdot \mathbf{v})$ represents the inner product between the two vectors $\mathbf{u}$ and $\mathbf{v}$. Actually, the SVM yields support vectors $\mathbf{c}_i^*$ ($i = 1, \dots, M$) from the training data, and the resultant output function of linear SVM for an input vector $\mathbf{c}$ is given by

$$f(\mathbf{c}) = \sum_{i=1}^M \alpha_i^* z_i (\mathbf{c}_i^*, \mathbf{c}) + b \quad (2)$$

where α_i^* is the solution of the quadratic programming problem. In addition, in order to generalize the target input space, the output function on the nonlinear SVM incorporating the kernel function is finally given for

$$f(\mathbf{c}) = \sum_{i=1}^M \alpha_i^* z_i K(\mathbf{c}_i^*, \mathbf{c}) + b \quad \begin{matrix} \text{male} \\ \geq \\ \text{female} \end{matrix} \quad \eta \quad (3)$$

where K denotes the applied kernel function and η is the threshold for gender identification. This time, comparing $f(\mathbf{c})$ and a given threshold value reveals the SVM-based gender identification decision statistic.

3 MFCC weight optimization using MCE training

One interesting insight pertaining to the MFCC-based SVM scheme for gender identification is that no weights are assigned to the MFCC [6, 7]. To incorporate the different contributions of the MFCC in discriminating gender from human voice, we propose a technique adopting various weights for the MFCC such as $w_k c_k$. Here, the weights $\{w_k\}$ should satisfy the following conditions:

$$\sum_{k=1}^N w_k = 1, \quad (4)$$

$$w_k \geq 0. \quad (5)$$

Let $\mathbf{c}_w = \{w_1 c_1, w_2 c_2, \dots, w_N c_N\}$ represent the optimally weighted MFCC vector. Two discriminant functions, g_m and g_f , are given for deciding whether each frame of a given speech signal belongs to male or female as follows:

$$g_m(\mathbf{c}_w) = f(\mathbf{c}_w) - \theta, \quad (6)$$

$$g_f(\mathbf{c}_w) = \theta - f(\mathbf{c}_w) \quad (7)$$

where θ denotes a threshold value of the combined score. As the next step, estimation of the weights is conducted based on the discriminative weight training framework in which the generalized probabilistic descent (GPD) technique is adopted. Let $\mathcal{D}(\mathbf{c}_w(t))$ denote the misclassification degree of the training data $\mathbf{c}_w(t)$ at time t . Then

$$\mathcal{D}(\mathbf{c}_w(t)) = \begin{cases} -g_m(\mathbf{c}_w(t)) + g_f(\mathbf{c}_w(t)) & \text{if true class is male} \\ -g_f(\mathbf{c}_w(t)) + g_m(\mathbf{c}_w(t)) & \text{if true class is female} \end{cases} \quad (8)$$

where the classification is considered to be right when (8) is negative. Here, the GPD method approximates the empirical classification error by a smooth objective function, which is the 0–1 step loss function defined by

$$L(t) = \frac{1}{1 + \exp(-\gamma \mathcal{D}(\mathbf{c}_w(t)))}, \quad \gamma > 0 \quad (9)$$

where γ denotes the gradient of the sigmoid function. Once the parameter γ is specified, the weights are substantially trained such that

$$\{\hat{w}_k\} = \arg \min_{\{w_k\}} L. \quad (10)$$

It is known that the steepest descent method is a predominant way to optimize the weights according to the above rule. Note that direct application of the steepest descent technique can not be allowed due to the constraints on the weights as specified in (5). Hence, we consider the following parameter transformation

$$\tilde{w}_k = \log w_k \quad (k = 1, \dots, N). \quad (11)$$

Let $\{\tilde{w}_k(t)\}$ denote the set of estimates for the transformed weights at time t . This is then updated based on the steepest descent technique as follows:

$$\tilde{w}_k(t+1) = \tilde{w}_k(t) - \epsilon \left. \frac{\partial L(t)}{\partial \tilde{w}_k} \right|_{\tilde{w}_k = \tilde{w}_k(t)} \quad (12)$$

where $\epsilon(> 0)$ is the step size [8, 9]. After $\tilde{w}_k$ is updated, $\tilde{w}_k$ is inversely transformed to w_k using the following rule:

$$w_k = \frac{\exp(\tilde{w}_k)}{\sum_{i=1}^N \exp(\tilde{w}_i)} \quad (13)$$

where (13) includes normalization of the weights to satisfy the constraint in (4). As a result, fig. 1 corresponds to the derived optimal weights for the 13th-order MFCC in our experiment. By comparing $f(\mathbf{c}_w)$ and a given threshold value, a final decision rule based on the SVM is given as follows:

$$f(\mathbf{c}_w) = \sum_{i=1}^M \alpha_i^* z_i(\mathbf{c}_i^*, \mathbf{c}_w) + b \quad \begin{matrix} \text{male} \\ \geq \\ \text{female} \end{matrix} \quad \eta. \quad (14)$$

Fig. 1. Optimal weights based on MCE method.

Fig. 2. Receiver operating characteristics for gender identification with (proposed) and without [6] weight (a) male (b) female.

4 Experimental results

The performance of the proposed method was evaluated on speech data extracted from the OGI database, where each file has an average length of 5 s [10]. For training, we used 1000 speech files, which were produced by 34 male and 34 female speakers. Speech features for gender identification consisted of 12 cepstral coefficients and log energy. Each frame with 25 ms of the windowed signal was shifted by 10 ms. Especially for the SVM training, the radial basis function (RBF) kernel was experimentally chosen for the efficient nonlinear classification of the sparse input data. After the SVM training, we conducted the MCE scheme in which the parameters for defining the objective function L were selected such that $\gamma = 1$ and the step size for parameter update was set to $\epsilon = 1 - \frac{t}{40000}$. These values resulted in a smooth convergence of the objective function to the minimal point. Also, the threshold value θ of the combined score was set to 0 as the experimentally

chosen boundary in the middle of $f(c)$ from male speech and female speech.

For testing, we used 1000 different speech files spoken by 34 male and 34 female speakers from the well-known OGI database [7]. In Fig. 2, the receiver operating characteristic (ROC) curves, which show the trade-off between the reduction of P_{fa} (false-alarm probability) and the increase in detection probability (P_d), are displayed for all test files. From the results, we can see that the proposed algorithm outperformed the previous method [6] in discriminating gender at the relevant intervals. The test results confirm that the optimally weighted MFCC improves the performance of gender identification in actual conditions.

5 Conclusions

In this paper we have proposed a novel gender identification technique based on discriminative weight training in which optimally weighted MFCCs are integrated into the SVM for a robust gender identification. The proposed approach yields better performance than the conventional method.

Acknowledgments

This work was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute for Information Technology Advancement) (IITA-2008-C1090-0902-0010) and was financially supported by the Ministry of Knowledge Economy (MKE) and Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Strategic Technology.