MINIMUM CLASSIFICATION ERROR INTERACTIVE TRAINING FOR SPEAKER IDENTIFICATION

Yusuke Kida[†], *Hiroyoshi Yamamoto*[‡], *Chiyomi Miyajima*[‡†], *Keiichi Tokuda*[‡], *and Tadashi Kitamura*[‡]

† Graduate School of Informatics,
Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
‡ Graduate School of Engineering,
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
‡† Graduate School of Information Science
Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

ABSTRACT

This paper describes an online discriminative training algorithm aiming at achieving speaker identification on interactive robots. A robot incrementally acquires speakers' voice characteristics during the interaction with the speakers. We simulate the situation that the speakers never give their IDs and the robot can only know whether the identification decision was correct or not from the speaker's positive or negative behavioral reaction. The speaker models are adjusted based on this limited information using minimum classification error (MCE) training consisting of positive and negative adaptation. In cases of correct identification, the conventional MCE training algorithm can be used. We compare three kinds of negative adaptation algorithms for the cases of incorrect identification. Experimental results show that the combination of the positive and negative adaptation achieves faster convergence, and negative adaptation which adjusts only a misclassified speaker model reaches an identification rate of 80% four times faster than the positive adaptation alone.

1. INTRODUCTION

Speaker identification is a well-known person authentication technique and has wide application. Speaker recognition by interactive robots such as humanoid or pet robots is also expected to be realized. A similar idea was reported in [1], where faces images was used for identification.

In conventional speaker identification systems, speakers are enrolled in advance. However, human-friendly speaker enrollment is desirable for the interactive robots. In such cases, we need to choose incremental training where an initial model is gradually adapted to each speaker in an online fashion. The information obtained via the interaction between human and robot can be used for model adaptation. However, people generally do not tell robots much information about their IDs while the interaction. Therefore, the robot can obtain a limited feedback information whether the identified speaker ID was correct or not by seeing or hearing the emotional or behavioral reactions of the speaker when the robot calling the name of the identified speaker. This simple feedback loop can be performed through simple interaction between human and robots. For example, this framework can be applied to pet robots which can distinguish their family members voice. Every time they speak to the robot, the robot can enhance its identification ability.

To simulate this interactive training framework, we adopt Gaussian mixture model (GMM) for speaker modeling which is widely used for text-independent speaker identification [2]. Minimum classification error (MCE) training is applied to the adaptive training of GMM parameters. The effectiveness of MCE training for speaker recognition has been reported in many previous works [3]–[7]. However, general MCE training requires the correct speaker ID of the input speech for training and it can not be simply applied to the supposed situation. The conventional MCE training can be applied only in cases of correct identification because the correct speaker ID can not be obtained in cases of incorrect identification. In this paper, we propose a new interactive training algorithm consisting of positive and negative adaptation which can be used without the information of speaker IDs. The effectiveness of the proposed method is evaluated in simulated experiments.

This paper is organized as follows. Section 2 presents the MCE training of GMM, and the proposed algorithms are described in Section 3. The experimental conditions and results are reported in Section 4, and conclusions and future works are given in Section 5.

2. MCE TRAINING FOR GMM SPEAKER MODEL

This section describes MCE training for GMM speaker model. Model parameters estimated by maximum likelihood do not guarantee to minimum classification error. Therefore, MCE training based on the generalized probabilistic descent (GPD) method [8] is applied to the parameters of GMM.

2.1. Definition of Loss Function

For the MCE training, the misclassification measure of training data $\boldsymbol{X}_k = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ for speaker k is defined as

$$d_k(\boldsymbol{X}_k; \boldsymbol{\Theta}) = -g_k(\boldsymbol{X}_k; \boldsymbol{\Theta}) + G_k(\boldsymbol{X}_k; \boldsymbol{\Theta}), \quad (1)$$

$$G_k(\boldsymbol{X}_k; \boldsymbol{\Theta}) = \log \left[\frac{1}{K-1} \sum_{j \neq k} \exp\{g_j(\boldsymbol{X}_k; \boldsymbol{\Theta})\eta\} \right]^{\frac{1}{\eta}}. \quad (2)$$

These can be written as follows when infinity is substituted for η which controls comparing operation:

$$d_k(\boldsymbol{X_k}; \boldsymbol{\Theta}) = -g_k(\boldsymbol{X_k}; \boldsymbol{\Theta}) + \max_{y \neq k} g_y(\boldsymbol{X_k}; \boldsymbol{\Theta}), \quad (3)$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k\}$ denotes the speaker model parameter set of GMM, and $g_k(\cdot;\cdot)$ is defined by the log likelihood of $\boldsymbol{X}_k$ for speaker model $\boldsymbol{\theta}_k$. The loss function is defined as a differentiable sigmoid function approximating the 0-1 step loss function:

$$l_k(\boldsymbol{X}_k;\, \boldsymbol{\theta}) = \big(1 + \exp(-\gamma \cdot d_k)\big)^{-1}, \quad (4)$$

where γ denotes the gradient of the sigmoid function. The goal of the discriminative training is to minimize the loss function based on the probabilistic descent method.

2.2. Parameter Adjustment of GMM

During the parameter training in the MCE training, the constraints of the GMM parameters, e.g., $p_m > 0$, should be satisfied. Hence, the GMM parameter set $\boldsymbol{\Theta}$ is transformed into a new model parameter set $\tilde{\boldsymbol{\Theta}}$.

$$\tilde{\boldsymbol{\Theta}} = \{\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_K\}, \quad (5)$$

$$\tilde{\boldsymbol{\theta}} = \{\tilde{p}_m, \tilde{\boldsymbol{\mu}}_m, \tilde{\boldsymbol{\Sigma}}_m, \mid m = 1, 2, \ldots, M\}, \quad (6)$$

where $\tilde{p}_m = \log c_m$, $\tilde{\mu}_{md} = \frac{\mu_{md}}{\Sigma_{mdd}}$,$\tilde{\Sigma}_{mdd} = \log \Sigma_{mdd}$. $\tilde{\boldsymbol{\Theta}}$ is updated at each iteration r as

$$\tilde{\boldsymbol{\Theta}}(r+1) = \tilde{\boldsymbol{\Theta}}(r) - \varepsilon_r \nabla l_k(\boldsymbol{X}_k;\, \tilde{\boldsymbol{\theta}}), \quad (7)$$

where ε_r is a monotonically decreasing learning step size at the r-th iteration. In this paper, $\tilde{\boldsymbol{\Theta}}$ is sequentially adjusted every time a training sample $\boldsymbol{X}_k$ is given (i.e., sample-by-sample mode).

The gradient of (7) is obtained as follows.

$$\nabla_{\tilde{\boldsymbol{\theta}}_j} l_k(\boldsymbol{X}_k; \tilde{\boldsymbol{\theta}}) = \frac{\partial l_k}{\partial d_k} \frac{\partial d_k}{\partial g_j} \cdot \nabla_{\tilde{\boldsymbol{\theta}}_j} g_j(\boldsymbol{X}_k; \tilde{\boldsymbol{\theta}}), \quad (8)$$

where $\frac{\partial l_k}{\partial d_k}$, $\frac{\partial d_k}{\partial g_j}$, $\nabla_{\tilde{\boldsymbol{\theta}}_j} g_j(\boldsymbol{X}_k; \tilde{\boldsymbol{\theta}})$ are given by

$$\frac{\partial l_k}{\partial d_k} = \gamma l_k (1 - l_k), \quad (9)$$

$$\frac{\partial d_k}{\partial g_j} = \begin{cases} -1 & j = k \\ 1 & j = y \\ 0 & \text{otherwise} \end{cases}, \quad (10)$$

$$\nabla_{\tilde{\boldsymbol{\theta}}_j} g_j(\boldsymbol{X}; \tilde{\boldsymbol{\theta}}) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{b_j(\boldsymbol{x}_t)} \nabla_{\tilde{\boldsymbol{\theta}}_j} b_j(\boldsymbol{x}_t). \quad (11)$$

The gradient of $b_y(\boldsymbol{x}_t)$ with respect to each element in $\tilde{\boldsymbol{\theta}}_j$ is obtained by the following formulae, where the subscript j is dropped for the simplicity of notation.

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{p}_m} = p_m \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_m \boldsymbol{\Sigma}_m), \quad (12)$$

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{\mu}_{md}} = \frac{x_{td} - \mu_{md}}{\sigma_{md}} p_m \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_m \boldsymbol{\Sigma}_m), \quad (13)$$

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{\Sigma}_{md}} = \left\{ \left(\frac{x_{td} - \mu_{md}}{\sigma_{md}} \right)^2 - 1 \right\} p_m \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_m \boldsymbol{\Sigma}_m). \quad (14)$$

3. PROPOSED NEGATIVE TRAINING ALGORITHMS BASED ON MCE

In this section, three kinds of training algorithms are proposed. For our simulated situation, speakers' IDs can not be obtained. Hence, general MCE training can be applied only in cases of correct identification. Therefore, we propose new training algorithms for the cases of incorrect identification based on MCE training.

3.1. Proposed Negative Training Algorithm A

Speakers' IDs are referred to by two equations of the misclassification measure d_k and $\frac{\partial d_k}{\partial g_j}$ in MCE training. The two equations are needed to be redefined for negative training.

For proposed training algorithm A, the misclassification measure of training data X_k for speaker k is defined as follows:

$$d_n(\boldsymbol{X}_k; \boldsymbol{\Theta}) = -\max_{p \neq n} g_p(\boldsymbol{X_k}; \boldsymbol{\Theta}) + g_n(\boldsymbol{X}_k; \boldsymbol{\Theta}), \quad (15)$$

where n denotes the speaker which is the result of identification. $\frac{\partial d_n}{\partial g_j}$ is given as follows:

$$\frac{\partial d_n}{\partial g_j} = \begin{cases} 1 & j = n \\ 0 & j \neq n \end{cases}. \quad (16)$$

Updating speaker models using proposed algorithms is shown in Fig.1. There are three speaker models a, b and c, each of which is shown of a circle. The cross mark corresponds to the input speech from speaker a. Positive adaptation is shown in (Pos). Assume that the nearest speaker model for the input speech is a. In this case speaker a is chosen as the identification result. In this

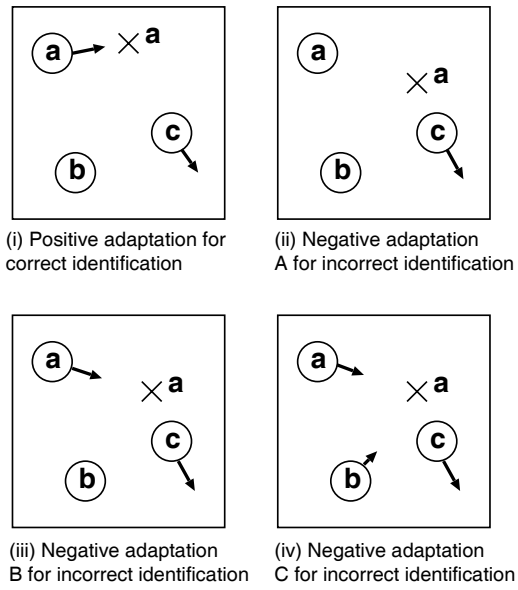

Fig. 1. Speaker model adjustment in proposed algorithms.

correct identification case, model a is updated toward the input speech and the speaker c which has the second-best likelihood is updated toward the opposite direction from it.

Negative adaptation using algorithm A is shown as (Neg A). In this case, the input speech of a is near to model c. Therefore, c is chosen as the result. However this is not correct. Accordingly, the misclassified model c is updated to the opposite direction from the speech input. These are the negative adaptation for algorithm A.

3.2. Proposed Negative Training Algorithm B

For proposed training algorithm B, the misclassification measure and $\frac{\partial d_n}{\partial g_j}$ is given as follows:

$$d_n(\boldsymbol{X}_k;\boldsymbol{\Theta}) = -\max_{p\neq n} g_p(\boldsymbol{X_k};\boldsymbol{\Theta}) + g_n(\boldsymbol{X}_k;\boldsymbol{\Theta}), \quad (17)$$

$$\frac{\partial d_n}{\partial g_j} = \begin{cases} 1 & j = n \\ -1 & j = p \\ 0 & \text{otherwise} \end{cases} . \quad (18)$$

Updating speaker models using algorithm B is shown as (Neg B) in Fig.1. Positive adaptation is the same way as in algorithm A. In cases of incorrect identification, second-best likelihood speaker a has the highest possibility to be the correct speaker. Therefore, algorithm B assumes a to be a correct speaker, and a is updated toward the speech input. In addition, model c is updated in the same way as in algorithm A.

3.3. Proposed Negative Training Algorithm C

Algorithm C defines the misclassification measure as follows:

$$d_n(\boldsymbol{X}_k;\boldsymbol{\Theta}) = -G_n(\boldsymbol{X}_k;\boldsymbol{\Theta}) + g_n(\boldsymbol{X}_k;\boldsymbol{\Theta}). \quad (19)$$

$\frac{\partial d_n}{\partial g_j}$ is given as follows:

$$\frac{\partial d_n}{\partial g_j} = \begin{cases} 1 & j = n \\ -\dfrac{\exp\{\eta g_j(\boldsymbol{X}_k;\Lambda)\}}{\sum_{i,i\neq n}\exp\{\eta g_i(\boldsymbol{X}_k;\Lambda)\}} & j \neq n \end{cases} . \quad (20)$$

Updating speaker models using algorithm C is shown as (Neg C). The misclassified model c is updated in the same way as the other algorithms. However, algorithm C assumes all the speakers except for c to be a correct speaker, and updates all the speaker models according to their likelihood for flexible updating of speaker models. Model c is updated the most, and the least is b. For algorithm C, less likelihood speaker model is updated less.

4. EXPERIMENTAL EVALUATION

4.1. Database and Experimental Conditions

An online text-independent speaker identification experiment was conducted to evaluate the adaptation algorithms using ten speakers in the ATR Japanese speech database.

The speech data was down-sampled from 20kHz to 10kHz, windowed at a 10-ms frame rate using a 25.6-ms Blackman window, and parameterized into 12 mel-cepstral coefficients excluding zero-th coefficient with a mel-cepstral analysis technique.

A 32-component GMM with diagonal covariance matrices was used for modeling each speaker. The GMM parameters were initialized with the EM algorithm using ten words randomly chosen from 100 words uttered by other ten male speakers. The speakers used for evaluation were not included in the data used for initialization. 216 words per speaker (2160 words in total) were used three times for adaptation and the total number of adaptation steps was 6480. All words were completely shuffled and used one by one. After every 100 word adaptation, the identification ability of the system was evaluated using 5200 words (520 words for each speaker) not used for adaptation.

In this experiment, as a first approach to the online speaker identification, the number of speakers who might talk to the robot was given in advance. However, automatic estimation of the number of target persons is a difficult problem for online person identification [9] and is also one of our future works.

4.2. Results

Figure 2 compares the following adaptation algorithms:

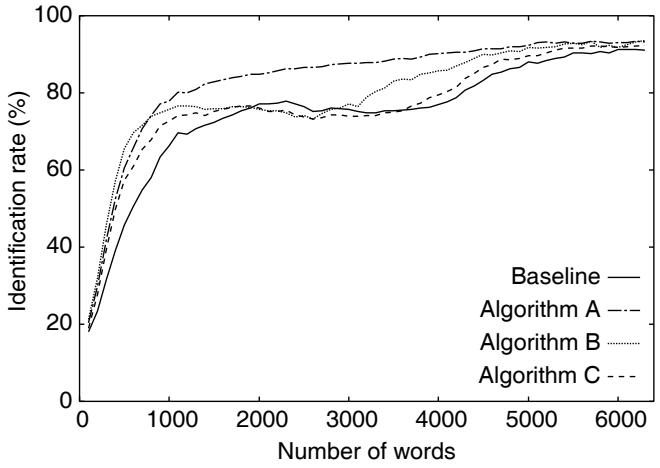

Fig. 2. Comparison of accuracy between the proposed adaptation algorithms.

Baseline: Positive adaptation alone,

Algorithm A: Positive and Negative A adaptation,

Algorithm B: Positive and Negative B adaptation,

Algorithm C: Positive and Negative C adaptation.

The horizontal axis corresponds to the number of words used for adaptation. The vertical axis corresponds to the identification accuracy evaluated using the 5200 words. The identification rate at an earlier phase of adaptation was quite low because the model parameters were initialized using randomly chosen speech data from different speaker data sets. The identification rates of all the adaptation methods converged to about 93% at the end of adaptation. The difference between these methods was the speed of adaptation. All the proposed algorithms A, B and C achieved faster adaptation than the baseline method, especially in an early stage of the adaptation. The number of words used for adaptation until the identification rate exceeded 80% were 4300, 1100, 3300 and 4100, for the baseline method, algorithms A, B and C, respectively. Algorithm A reached 80% accuracy four times faster than the baseline method. Algorithm A showed the fastest adaptation and the most stable performance. We expect that the adaptation speed can be much faster if we tune the learning step size for adaptation. Finding an appropriate learning step size is one of the issues we would like to address in the future.

Algorithms B and C used hypotheses for negative adaptation and updated speaker models besides the misclassified speaker model, while algorithm A did not use hypothesis and updated only the misclassified speaker model. The experiment showed that the simplest model adjustment without hypothesis was most effective for the negative adaptation.

5. CONCLUSIONS

This paper proposed an interactive training algorithm of speaker models for speaker identification by robots, which used positive and negative adaptation based on minimum classification error criterion. We conducted a speaker identification experiment and compared three kinds of negative adaptation algorithms. We confirmed the effectiveness of the negative adaptation algorithms and the simplest algorithm that updated only a misclassified speaker model showed the fastest convergence of adaptation and the most stable performance.

Our future works include further study of effective adaptation methods and the integration of face images and speech for the person identification on interactive robots.

6. REFERENCES

[1] L. Aryananda, "Online and unsupervised face recognition for humanoid robot: toward relationship with people," *International Conference on Humanoid Robots*, 2001.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Process.*, vol.3, no.1, pp.72–83, Jan. 1995.

[3] C.-S. Liu, C.-H. Lee, B.-H. Juang, and A. E. Rosenberg, "A study on minimum error discriminative training for speaker recognition," *J. Acoust. Soc. Am.*, vol.97, no.1, pp.637–648, Jan. 1995.

[4] O. Siohan, A. E. Rosenberg, and S. Parthasarathy, "Speaker identification using minimum classification error training," *ICASSP-98*, vol.1, pp.109–112, May 1998.

[5] C. Miyajima, K. Tokuda, and T. Kitamura, "Minimum classification error training for speaker identification using Gaussian mixture models based on multi-space probability distribution," *EUROSPEECH-2001*, vol.4, pp.2837–2840, Sept. 2001.

[6] F. Korkmazskiy and B.-H. Juang, "Discriminative adaptation for speaker verification," *ICSLP-96*, vol.3, pp.1744–1747, Oct. 1996.

[7] C. Martin del Alamo, J. lvarez, C. de la Torre, F. J. Payotos, and L. Hernndez, "Incremental speaker adaptation with minimum error discriminative training for speaker identification," *ICSLP-96*, vol.3, pp.312–315, Oct. 1996.

[8] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Process.*, vol.40, no.12, pp.3043–3054, Dec. 1992.

[9] D. Liu and F. Kubala, "Online speaker clustering," *ICASSP-2004*, vol.1 pp.333–336, May 2004.