Minimum Classification Error Training for Speaker Identification Using Gaussian Mixture Models Based on Multi-Space Probability Distribution

Chiyomi Miyajima, Keiichi Tokuda, and Tadashi Kitamura

Department of Computer Science, Nagoya Institute of Technology, Nagoya 466-8555, Japan
{chiyomi, tokuda, kitamura}@ics.nitech.ac.jp

Abstract

In our previous work, we have proposed a speaker modeling technique using spectral and pitch features for text-independent speaker identification based on Multi-Space Probability Distribution Gaussian Mixture Models (MSD-GMMs). We have presented a maximum likelihood (ML) estimation procedure for the MSD-GMM parameters and demonstrated its high recognition performance. In this paper, we describe a minimum classification error (MCE) training approach based on the generalized probabilistic descent (GPD) method [9] for text-independent speaker identification using MSD-GMMs. In addition, we introduce stream weighting parameters into the MSD-GMM speaker models to improve the recognition ability by appropriately weighting the spectral and pitch streams. Several works have demonstrated the utility of MCE training in speaker recognition [10], and MCE training has also been successfully applied to stream weight estimation instead of ML estimation [11]. In this study, the stream weights as well as MSD-GMM parameters are estimated using MCE training. The MCE-based MSD-GMM speaker models are evaluated for an 80 speaker text-independent speaker identification task.

In Section 2 we introduce stream weights into the MSD-GMM and describe a speaker modeling technique using spectral and pitch features. Section 3 presents the MCE training procedure for the MSD-GMM parameters and stream weights. Section 4 reports experimental results, and Section 5 gives conclusions and future works.

1. Introduction

Prosodic features such as pitch features as well as spectral features contain much speaker individuality [1, 2]. Several works have reported that speaker recognition performance can be improved by the use of pitch features along with spectral features [3, 4, 5, 6, 7]. Conventional approaches to integrating spectral and pitch features are i) two separate models are trained for spectral and pitch features, respectively, and their scores are combined in a recognition stage [3, 4], ii) two separate models for voiced and unvoiced parts are trained, respectively, and their scores are combined [5, 6]. For example, two Gaussian Mixture Models (GMMs) were used for voiced and unvoiced parts in [6], where the input observations were concatenations of cepstral coefficients and log fundamental frequency (log $F_0$) for voiced frames, and cepstral coefficients alone for unvoiced frames, i.e., $(D+1)$-dimensional vectors for voiced parts and $D$-dimensional vectors for unvoiced parts. These two kinds of vectors having different dimensionalities require their respective GMMs because the probability distribution of the traditional GMM [8] is defined on a single vector space.

In contrast to the above two approaches, we jointly model spectral and pitch features of both voiced and unvoiced parts exploiting Gaussian Mixture Models based on the Multi-Space Probability Distribution (MSD-GMM) [7]. The MSD-GMM allows us to model feature vectors having variable dimensionality including zero-dimensional vectors, i.e., discrete symbols. Consequently, continuous pitch values of voiced frames and discrete symbols representing “unvoiced frame” can be modeled using an MSD-GMM in a unified framework, and spectral and pitch features are jointly modeled by a multi-stream MSD-GMM, i.e., each speaker is modeled by a single statistical model. In [7], we have presented maximum likelihood (ML) estimation formulae for the MSD-GMM parameters and demonstrated its high recognition performance.

In this paper, we describe a minimum classification error (MCE) training approach based on the generalized probabilistic descent (GPD) method [9] for text-independent speaker identification using MSD-GMMs. In addition, we introduce stream weighting parameters into the MSD-GMM speaker models to improve the recognition ability by appropriately weighting the spectral and pitch streams. Several works have demonstrated the utility of MCE training in speaker recognition [10], and MCE training has also been successfully applied to stream weight estimation instead of ML estimation [11]. In this study, the stream weights as well as MSD-GMM parameters are estimated using MCE training. The MCE-based MSD-GMM speaker models are evaluated for an 80 speaker text-independent speaker identification task.

2. Multi-Stream MSD-GMM

2.1. Likelihood Calculation for MSD-GMM

Let us assume that a given observation at time $t$ $o_t$ consists of $S$ information sources (streams). The $s$-th stream $o_{ts}$ has a set of space indices $X_{ts}$, and an observation vector of variable dimensionality $x_{ts}$, that is

$$o_{ts} = (o_{t1}, o_{t2}, \ldots, o_{tS}),$$

where $x_{ts}$ is assumed to be an observation vector from one of the vector spaces represented by the indices in $X_{ts}$, and $X_{ts}$ is a nonempty subset of all possible space indices for the $s$-th stream:

$$X_{ts} \subseteq \{1, 2, \ldots, G_s\}.$$ 

All the spaces in $X_{ts}$ have the same dimensionality as that of $x_{ts}$. The output distribution of $o_{ts}$ for $S$-stream MSD-GMM is defined as

$$b(o_{ts} \mid \lambda) = \sum_{m=1}^{M} c_m \prod_{s=1}^{S} p_{ms}(o_{ts}),$$

where $c_m$ is mixture weight for the $m$-th mixture component. The observation probability of $o_{ts}$ for mixture $m$ is given by the multi-space probability distribution (MSD):

$$p_{ms}(o_{ts}) = \sum_{g \in X_{ts}} w_{mg} N_{msg}^{D_{sg}}(x_{ts}),$$

where $w_{mg}$ is space weight for the $g$-th vector space of the $s$-th stream and $N_{msg}^{D_{sg}}(\cdot)$ is the $D_{sg}$-variate Gaussian function with
mean vector $\mu_{msg}$ and covariance matrix $\Sigma_{msg}$ for the case $D_{sg} > 0$. For simplicity of notation, we define $N_{msg}(\cdot) \equiv 1$ for the case $D_{sg} = 0$. Note here that the multi-space probability distribution is equivalent to continuous probability distribution when $D_{sg} = 0$. MSD-GMM can be assumed to be a generalized GMM, which includes the traditional GMM as a special case when $S = 1$, $G_1 = 1$, and $D_{11} > 0$.

For an observation sequence
\[ O = (o_1, o_2, \ldots, o_T), \]  
the likelihood of MSD-GMM $\lambda$ is given by
\[ P(O \mid \lambda) = \prod_{t=1}^{T} b(o_t \mid \lambda). \]

2.2. Stream Weighting

Equation (4) can be slightly modified to control a weighting factor for each stream by introducing a stream weight parameter $\gamma_{msg}$, that is
\[ b(o_t \mid \lambda) = \sum_{m=1}^{M} \sum_{s=1}^{S} c_m \cdot \mathbb{1}_{\{p_{ms}(o_{ts})\}} \gamma_{msg}. \]

Note that $b(o_t \mid \lambda)$ does not satisfy the stochastic restrictions.

Figure 1 illustrates an example of the $m$-th mixture component of a three-stream MSD-GMM ($S = 3$). The sample space of the first stream consists of four spaces ($G_1 = 4$), among which, the second and the third spaces are triggered by the space indices and $p_{11}(o_{t1})$ becomes the sum of the two weighted Gaussians. The second stream has only one space ($G_2 = 1$) and always outputs its Gaussian as $p_{21}(o_{t2})$. The third stream consists of two spaces ($G_3 = 2$), where a zero-dimensional space is selected, and its space weight $w_{m32}$ (a discrete probability) becomes $p_{m3}(o_{t3})$.

2.3. Speaker Modeling Based on MSD-GMM

Spectral features are generally represented by $D$-dimensional vectors of cepstral coefficients with continuous values ($D$ is constant). On the other hand, pitch features are represented by one-dimensional continuous values of log fundamental frequency ($\log F_0$) for voiced frames and discrete symbols representing “unvoiced” for unvoiced frames because pitch values are defined only in voiced segments. Hence, each speaker can be modeled by a two-stream MSD-GMM ($S = 2$) (Fig. 2). The first stream is for the spectral feature and the second stream is for the pitch feature. The spectral stream has a $D$-dimensional vector space ($G_1 = 1$) and the pitch stream has two spaces (a one-dimensional space and a zero-dimensional space) for voiced and unvoiced parts ($G_2 = 2$).

3. MCE Training for MSD-GMM Parameters and Stream Weights

MCE training based on the GPD method optimizes the parameters of a classifier in a pattern recognizer using a gradient technique [9]. In this paper, we apply the GPD method to MSD-GMM based speaker identification. In this case, the adjustable parameter set $\lambda$ includes the entire parameters of $N$ MSD-GMM speaker models
\[ \lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}, \]
where $\lambda_n$ includes mixture weights, stream weights, space weights, means, and variances of the $n$-th speaker model.
where \(1(\mathcal{X})\) is the indicator function for a logical variable \(\mathcal{X}\) defined as
\[
1(\mathcal{X}) = \begin{cases} 
1, & \text{if } \mathcal{X} \text{ is true} \\
0, & \text{otherwise}
\end{cases}.
\]

### 3.2. Minimization of Empirical Loss

The parameter set \(\Lambda\) is sequentially adjusted every time a training sample \(O_t\) from speaker class \(C_k\) is given according to
\[
\Lambda^{(r+1)} = \Lambda^{(r)} - \epsilon^{(r)} U^{(r)} \nabla_{\Lambda} \ell_k(O_t; \Lambda) |_{\Lambda^{(r)}} ,
\]
where \(\epsilon^{(r)}\) is a monotonically decreasing learning step size and \(U^{(r)}\) is a positive definite learning matrix at the \(r\)-th iteration.

During the parameter adaptation, the constraints of the parameters, such as \(c_m > 0\) and \(\gamma_{ms} > 0\), should be satisfied, and means are normalized with variances. The following parameter transformations allow us to maintain these constraints:
\[
\tilde{c}_m = \log \epsilon_m, \quad \tilde{\gamma}_{ms} = \log \gamma_{ms},
\]
\[
\tilde{\mu}_{msgd} = \frac{\mu_{msgd}}{\sigma_{msgd}}, \quad \tilde{\sigma}_{msgd}^2 = \log \sigma_{msgd}^2 ,
\]
where \(\mu_{msgd}\) is the \(d\)-th element of \(\mu_{msgd}\) and \(\sigma_{msgd}\) is the \(d\)-th diagonal element of (diagonal) matrix \(\Sigma_{msgd}\). Using the new parameter set \(\Lambda\), (16) is rewritten as
\[
\tilde{\Lambda}^{(r+1)} = \tilde{\Lambda}^{(r)} - \tilde{\epsilon}^{(r)} U^{(r)} \nabla_{\tilde{\Lambda}} \ell_k(O_t; \tilde{\Lambda}) |_{\tilde{\Lambda}^{(r)}} ,
\]
The parameter sequence produced by (22) (with probability one) to the minimum point of (14) for large \(H\) [9].

According to the chain rule, the gradient \(\nabla_{\tilde{\Lambda}} \ell_k\) (the subscript \(h\) is omitted and \(\ell_k(O_t; \tilde{\Lambda})\) is shortened to \(\ell_k\) to simplify the notation) in (22) can be rewritten as
\[
\nabla_{\tilde{\Lambda}} \ell_k = \frac{\partial \ell_k}{\partial \tilde{\Lambda}} \sum_{n=1}^N \sum_{i=1}^{T} \frac{\partial \mu_{n}}{\partial \tilde{b}(\omega_i | \lambda_n)} \nabla \tilde{b}(\omega_i | \lambda_n),
\]
where
\[
\frac{\partial \ell_k}{\partial \tilde{c}_m} = \alpha \ell_k (1 - \ell_k),
\]
\[
\frac{\partial \ell_k}{\partial \gamma_{ms}} = \begin{cases} 
-1, & n = k \\
1, & n = \arg \max_i g_i \\
0, & \text{otherwise}
\end{cases},
\]
\[
\frac{\partial \mu_{n}}{\partial \tilde{b}(\omega_i | \lambda_n)} = \frac{1}{T \tilde{b}(\omega_i | \lambda_n)},
\]
Dropping the subscripts \(t\) and \(n\) for simplicity of notation, each component of \(\nabla_{\tilde{\Lambda}} \ell_k(\omega_i | \lambda)\) can be derived as follows:
\[
\frac{\partial \tilde{b}(\omega_i | \lambda)}{\partial \tilde{c}_m} = \xi_{m}(\omega_i),
\]
\[
\frac{\partial \tilde{b}(\omega_i | \lambda)}{\partial \gamma_{ms}} = \gamma_{ms} \log \{p_{m}(\omega_i)\} \xi_{m}(\omega_i).
\]
For \(g \in X_s\),
\[
\frac{\partial \tilde{b}(\omega_i | \lambda)}{\partial \tilde{w}_{msgd}} = \xi_{msgd}(\omega_i),
\]
\[
\frac{\partial \tilde{b}(\omega_i | \lambda)}{\partial \tilde{\mu}_{msgd}} = \frac{\tilde{x}_{sd} - \tilde{\mu}_{msgd}}{\tilde{\sigma}_{msgd}} \xi_{msgd}(\omega_i),
\]
\[
\frac{\partial \tilde{b}(\omega_i | \lambda)}{\partial \tilde{\sigma}_{msgd}^2} = \left( \frac{\tilde{x}_{sd} - \tilde{\mu}_{msgd}}{\tilde{\sigma}_{msgd}} \right)^2 \xi_{msgd}(\omega_i).
\]

where \(x_{sd}\) is the \(d\)-th element of vector \(x_s\), and
\[
\xi_{m}(\omega_i) = c_m \prod_{s=1}^{S} \{p_{m}(\omega_s)\}^{\gamma_{ms}},
\]
\[
\xi_{msgd}(\omega_i) = \frac{\gamma_{msgd} \tilde{w}_{msgd} \tilde{\mu}_{msgd} (\omega_i)}{\tilde{p}_{ms}(\omega_i)} \xi_{m}(\omega_i).
\]

For \(g \notin X_s\),
\[
\frac{\partial \tilde{b}(\omega_i | \lambda)}{\partial \tilde{w}_{msgd}} = 0,
\]
\[
\frac{\partial \tilde{b}(\omega_i | \lambda)}{\partial \tilde{\mu}_{msgd}} = 0,
\]
\[
\frac{\partial \tilde{b}(\omega_i | \lambda)}{\partial \tilde{\sigma}_{msgd}^2} = 0.
\]

After the adjustment of \(\tilde{\Lambda}\) is completed, \(\tilde{\Lambda}\) is transformed back to \(\Lambda\) as follows:
\[
\tilde{c}_m = \frac{\exp \tilde{c}_m}{\sum_{l=1}^{M} \exp \tilde{c}_l},
\]
\[
\tilde{\gamma}_{ms} = \sum_{h=1}^{N} \exp \tilde{\gamma}_{ms},
\]
\[
\tilde{\mu}_{msgd} = \tilde{\mu}_{msgd} \tilde{\mu}_{msgd},
\]
\[
\tilde{\sigma}_{msgd}^2 = \exp \tilde{\sigma}_{msgd}^2 ,
\]

### 4. Experimental Evaluation

#### 4.1. Experimental Conditions

Text-independent speaker identification experiments were carried out for 80 speakers (40 males and 40 females) in the ATR Japanese speech database. Phonetically-balanced 216 words are used for training each speaker model, and 520 common words are used for testing. The number of tests was 41600 in total. Word boundaries were detected using log energy contours and silence parts at the beginning and end of the words were removed.

The speech data were down-sampled from 20 kHz to 10 kHz, windowed at a 10 ms frame rate with a 25.6 ms Blackman window, and parameterized into 13 mel-cepstral coefficients using a mel-cepstral estimation technique [13]. The 12 static parameters excluding the zero-th coefficient were used as spectral features. Fundamental frequency (\(F_0\)) was estimated using the RAPT method [14] with a 7.5 ms correlation window for every 10 ms. Pitch features were log \(F_0\) for the voiced frames and discrete symbols for unvoiced frames.

Each speaker was modeled by a GMM or a multi-stream MSD-GMM with diagonal covariance matrices. Baseline GMM and MSD-GMM parameters were ML-trained parameters initialized with an LBG codebook. Stream weights were initialized as 1.0. Identity matrices were used as learning matrices and the learning step size was initialized as \(\epsilon^{(0)} = 0.2\). The slope of the sigmoid function \(\alpha\) for each model set \(\Lambda\) was automatically chosen before the training using the variance of the misclassification measures for all the training samples \(v\), according to \(\alpha = 4 / \sqrt{27 \pi v}\). GPD training was iterated over 20 epochs, with the order of the given training samples being shuffled at the beginning of each training epoch.
6.20 GMM (32 mix.)
5.76 V-GMM (24 mix.) + UV-GMM (8 mix.)
5.22 MSD-GMM (32 mix.)

Figure 3: Comparison of MSD-GMM speaker models with conventional GMM speaker models.

4.2. Comparison with Conventional Speaker Models

We compared the MSD-GMM speaker identification system with three kinds of conventional systems. Figure 3 shows speaker identification error rates with 95% confidence intervals (CIs) when using 32 component ML-trained speaker models. In the figure, (a) corresponds to a conventional GMM speaker model using a spectral feature alone, (b) represents a speaker model consisting of two GMMs for spectral and pitch features, (c) is a speaker model consisting of two GMMs for voiced (V) and unvoiced (UV) parts [6], and (d) corresponds to the multi-stream MSD-GMM. For system (c), the optimum shares of the mixture components for the V-GMM and the UV-GMM were posteriorly tuned and best results were obtained with the ratio V:UV=3:1.

The additional use of pitch information improved system performance, and (b), (c) and (d) using both spectral and pitch information gave much better performance than the conventional GMM system using a spectral feature alone. Among the three systems, the MSD-GMM system gave the best results and achieved 16% error reduction over the GMM system.

4.3. Results for MCE Training

Figure 4 shows the results for MCE training. In the figure, baseline systems (A) and (C) are identical to the ML-based systems (a) and (d) in Fig. 3, respectively. The baseline systems (A) and (C) were improved to (B) and (E), respectively, by optimizing the model parameters. Results (D) and (F) were obtained after estimating the mixture-dependent stream weights based on MCE training for the systems (C) and (E), respectively. The error rate was significantly reduced by optimizing the MSD-GMM parameters based on MCE training (C →E). System performance was further improved by adjusting the stream weights for spectral and pitch streams and 16% error reduction was achieved over the ML-based MSD-GMM (C →F), resulting in 29% error reduction over the ML-based GMM (A →F).

5. Conclusions

This paper has described an MCE training procedure for text-independent speaker identification using MSD-GMMs. We introduced stream weights into the MSD-GMM to control the weighting factor for spectral and pitch streams. The stream weights as well as MSD-GMM parameters were automatically estimated based on the GPD method. The MCE training of MSD-GMM parameters significantly improved recognition ability and the optimal weighting of spectral and pitch streams achieved additional performance gains.

Further studies include investigation of stream weight initialization and tying strategies.

6. Acknowledgements

The authors would like to thank Prof. T. Kobayashi and Research Associate T. Masuko of Tokyo Institute of Technology for their valuable advice.

7. References