PAPER *Special Issue on Biometric Person Authentication*

Text-Independent Speaker Identification Using Gaussian Mixture Models Based on Multi-Space Probability Distribution

Chiyomi MIYAJIMA[†], ***Regular Member***, **Yosuke HATTORI**[†*], ***Nonmember***, **Keiichi TOKUDA**[†], **Takashi MASUKO**[††], **Takao KOBAYASHI**[††], ***and*** **Tadashi KITAMURA**[††], ***Regular Members***

SUMMARY This paper presents a new approach to modeling speech spectra and pitch for text-independent speaker identification using Gaussian mixture models based on multi-space probability distribution (MSD-GMM). MSD-GMM allows us to model continuous pitch values of voiced frames and discrete symbols for unvoiced frames in a unified framework. Spectral and pitch features are jointly modeled by a two-stream MSD-GMM. We derive maximum likelihood (ML) estimation formulae and minimum classification error (MCE) training procedure for MSD-GMM parameters. The MSD-GMM speaker models are evaluated for text-independent speaker identification tasks. The experimental results show that the MSD-GMM can efficiently model spectral and pitch features of each speaker and outperforms conventional speaker models. The results also demonstrate the utility of the MCE training of the MSD-GMM parameters and the robustness for the inter-session variability.

key words: *speaker identification, pitch, multi-space probability distribution, Gaussian mixture model, minimum classification error*

1. Introduction

There are many applications of text-independent speaker identification, including security control for restricted systems, speaker-adaptive speech recognition and automatic speaker labeling for audio indexing of recorded meetings. Intext-independent speaker identification, Gaussian mixture models (GMMs) and vector quantization techniques have been successfully applied to speaker modeling [1], [2]. Such identification systems mainly use spectral features represented by cepstral coefficients as speaker features.

Pitch features as well as spectral features contain much speaker specific information [3], [4]. However, most of speaker recognition studies in recent years have focused on using only spectral features. The main reasons for this are i) the use of pitch features alone could not give enough recognition performance and ii) pitch values are not defined in unvoiced segments and this complicates speaker modeling and feature integration.

Several works have reported that speaker recognition accuracy can be improved by the use of pitch features in addition to spectral features [5]–[8]. There are essentially two approaches to integrating spectral and pitch information: i) two separate models are used for spectral and pitch features and their scores are combined [5], [6], ii) two separate models for voiced and unvoiced parts are trained and their scores are combined [7], [8], e.g., two separate GMMs are used in [8], where the input observations are concatenations of cepstral coefficients and $\log F_0$ for voiced frames and cepstral coefficients alone for unvoiced frames. Since the probability distribution of the conventional GMM is defined on a single vector space, these two kinds of vectors require their respective models.

In this paper a new speaker modeling technique using a GMM based on multi-space probability distribution (MSD) [9], [10] is introduced. The GMM (MSD-GMM) [11] allows us to model feature vectors with 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. We derive maximum likelihood (ML) estimation formulae based on the expectation maximization (EM) algorithm and an minimum classification error (MCE) training procedure based on the generalized probabilistic descent (GPD) method [12], [13]. These speaker models are evaluated for text-independent speaker identification tasks and compared with conventional GMM speaker models.

The rest of the paper is organized as follows. In Sect. 2 we introduce a speaker modeling technique based on MSD-GMM. Sections 3 and 4 present the ML estimation and the GPD training procedures for MSD-GMM parameters, respectively. Section 5 reports experimental results, and Sect. 6 gives conclusions and future works.

Manuscript received September 19, 2000.
Manuscript revised January 24, 2001.

[†]The authors are with the Department of Computer Science, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.
[††]The authors are with the Department of Information Processing, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama-shi, 226–8502 Japan.
[*]Presently, with DENSO Corporation.

2. Multi-Stream MSD-GMM

2.1 Likelihood Calculation

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

$$\boldsymbol{o}_t = (\boldsymbol{o}_{t1}, \boldsymbol{o}_{t2}, \ldots, \boldsymbol{o}_{tS}), \tag{1}$$

$$\boldsymbol{o}_{ts} = (X_{ts}, \boldsymbol{x}_{ts}). \tag{2}$$

Note here that X_{ts} is a subset of all possible indices $\{1, 2, \ldots, G_s\}$, and all the spaces represented by the indices in X_{ts} have the same dimensionality as $\boldsymbol{x}_{ts}$.

We define the output probability distribution of an S-stream MSD-GMM λ for $\boldsymbol{o}_t$ as

$$b(\boldsymbol{o}_t \mid \lambda) = \sum_{m=1}^{M} c_m \prod_{s=1}^{S} p_{ms}(\boldsymbol{o}_{ts}), \tag{3}$$

where c_m is the mixture weight for the m-th mixture component. The observation probability of $\boldsymbol{o}_{ts}$ for mixture m is given by the multi-space probability distribution (MSD) [9], [10], that is

$$p_{ms}(\boldsymbol{o}_{ts}) = \sum_{g \in X_{ts}} w_{msg} \mathcal{N}^{D_{sg}}_{msg}(\boldsymbol{x}_{ts}), \tag{4}$$

where w_{msg} is the weight for the g-th vector space of the s-th stream and $\mathcal{N}^{D_{sg}}_{msg}(\cdot)$ is the D_{sg}-variate Gaussian function with mean vector $\boldsymbol{\mu}_{msg}$ and covariance matrix $\boldsymbol{\Sigma}_{msg}$ (for the case $D_{sg} > 0$). For simplicity of notation, we define $\mathcal{N}^0_{msg}(\cdot) \equiv 1$ (for the case $D_{sg} = 0$). Note here that the multi-space probability distribution (MSD) is equivalent to continuous probability distribution and discrete probability distribution when $D_{sg} \equiv n > 0$ and $D_{sg} \equiv 0$, respectively. Also, MSD-GMM is 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

$$\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T), \tag{5}$$

the likelihood of MSD-GMM λ is given by

$$P(\boldsymbol{O} \mid \lambda) = \prod_{t=1}^{T} b(\boldsymbol{o}_t \mid \lambda). \tag{6}$$

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 $p_{m1}(\boldsymbol{o}_{t1})$ becomes the sum of the two weighted Gaussians. The second stream has only one space ($G_2 = 1$) and always outputs its Gaussian as $p_{m2}(\boldsymbol{o}_{t2})$. 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 $p_{m3}(\boldsymbol{o}_{t3})$.

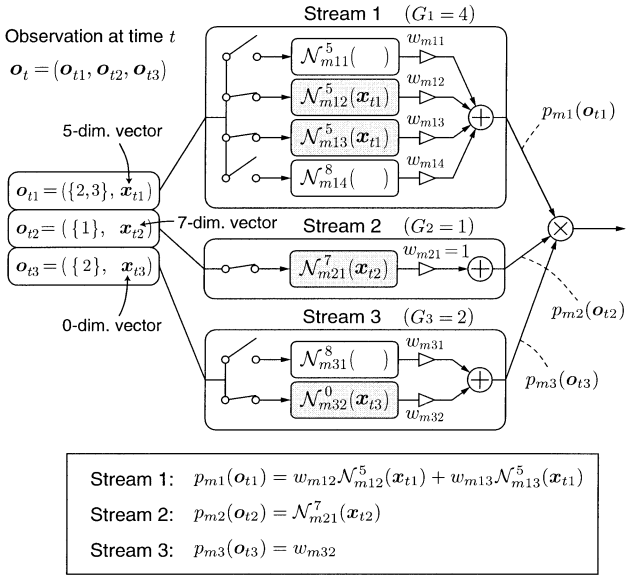


Fig. 1 An example of the m-th mixture component of a three-stream MSD-GMM.

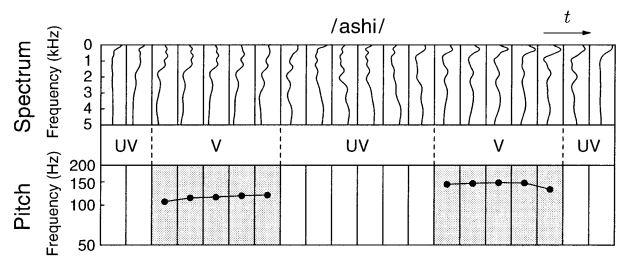


Fig. 2 An example of spectral and pitch sequences of the word /ashi/ spoken by a male speaker.

2.2 Speaker Modeling Based on MSD-GMM

Figure 2 shows an example of spectral and pitch sequences of the word /ashi/ spoken by a male speaker. Generally, spectral features are represented by multi-dimensional vectors of cepstral coefficients with continuous values. On the other hand, pitch features are represented by one-dimensional continuous values of log fundamental frequencies ($\log F_0$) in voiced frames and discrete symbols representing "unvoiced" in unvoiced frames because pitch values are defined only in voiced segments.

As shown in Fig. 3, each speaker can be modeled by a two-stream MSD-GMM ($S = 2$); the first stream is for the spectral feature and the second stream is for the pitch feature. The spectral stream has a D-dimensional space ($G_1 = 1$) and the pitch stream has two spaces (a

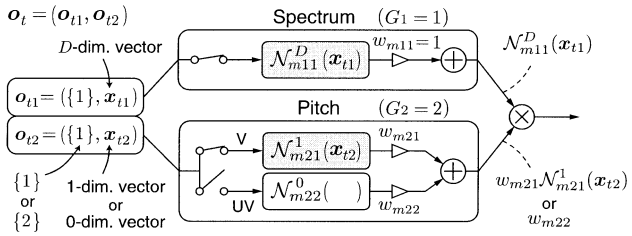


Fig. 3 The m-th mixture component in a two-stream MSD-GMM based on spectra and pitch.

one-dimensional space and a zero-dimensional space) for voiced and unvoiced parts ($G_2 = 2$).

3. ML-Estimation for MSD-GMM

In a similar way to the ML-estimation procedure in [10], $P(\boldsymbol{O} \mid \lambda)$ is increased by iterating the maximization of an auxiliary function $Q(\lambda', \lambda)$ over λ to improve current parameters λ' based on the EM algorithm.

3.1 Definition of Q-Function

The log-likelihood of λ for an observation sequence $\boldsymbol{O}$, a sequence of mixture components $\boldsymbol{i}$ and a sequence of space indices $\boldsymbol{l}$ can be written as

$$\log P(\boldsymbol{O}, \boldsymbol{i}, \boldsymbol{l} \mid \lambda) = \sum_{t=1}^{T} \log c_{i_t} + \sum_{t=1}^{T}\sum_{s=1}^{S} \log w_{i_t s l_{ts}} + \sum_{t=1}^{T}\sum_{s=1}^{S} \log \mathcal{N}^{D_{sl_{ts}}}_{i_t s l_{ts}}(\boldsymbol{x}_{ts}), \tag{7}$$

where

$$\boldsymbol{i} = (i_1, i_2, \ldots, i_T), \tag{8}$$
$$\boldsymbol{l} = (\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T), \tag{9}$$
$$\boldsymbol{l}_t = (l_{t1}, l_{t2}, \ldots, l_{tS}). \tag{10}$$

Hence the Q-function is defined as

$$\begin{aligned} Q(\lambda', \lambda) &= \sum_{\text{all } \boldsymbol{i}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{i}, \boldsymbol{l} \mid \lambda') \log P(\boldsymbol{O}, \boldsymbol{i}, \boldsymbol{l} \mid \lambda) \\ &= \sum_{\text{all } \boldsymbol{i}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{i}, \boldsymbol{l} \mid \lambda') \sum_{t=1}^{T} \log c_{i_t} \\ &\quad + \sum_{\text{all } \boldsymbol{i}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{i}, \boldsymbol{l} \mid \lambda') \sum_{t=1}^{T}\sum_{s=1}^{S} \log w_{i_t s l_{ts}} \\ &\quad + \sum_{\text{all } \boldsymbol{i}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{i}, \boldsymbol{l} \mid \lambda') \sum_{t=1}^{T}\sum_{s=1}^{S} \log \mathcal{N}^{D_{sl_{ts}}}_{i_t s l_{ts}}(\boldsymbol{x}_{ts}) \\ &= \sum_{m=1}^{M}\sum_{t=1}^{T} P(\boldsymbol{O}, i_t = m \mid \lambda') \log c_m \\ &\quad + \sum_{m=1}^{M}\sum_{s=1}^{S}\sum_{g=1}^{G_s}\sum_{t \in T(\boldsymbol{O}, s, g)} P(\boldsymbol{O}, i_t = m, l_{ts} = g \mid \lambda') \log w_{msg} \\ &\quad + \sum_{m=1}^{M}\sum_{s=1}^{S}\sum_{g=1}^{G_s}\sum_{t \in T(\boldsymbol{O}, s, g)} P(\boldsymbol{O}, i_t = m, l_{ts} = g \mid \lambda') \log \mathcal{N}^{D_{sg}}_{msg}(\boldsymbol{x}_{ts}), \end{aligned} \tag{11}$$

where

$$T(\boldsymbol{O}, s, g) = \{t \mid g \in X_{ts}\}. \tag{12}$$

3.2 Maximization of Q-Function

The first two terms of (11) have the form $\sum_{i=1}^{N} u_i \log y_i$, which attains a global maximum at the single point

$$y_i = \frac{u_i}{\sum_{j=1}^{N} u_j}, \quad \text{for} \quad i = 1, 2, \ldots, N, \tag{13}$$

under the constraints $\sum_{i=1}^{N} y_i = 1$ and $y_i \geq 0$. The maximization of the first term of (11) leads to the re-estimate of c_m:

$$\begin{aligned} c_m &= \frac{\sum_{t=1}^{T} P(\boldsymbol{O}, i_t = m \mid \lambda')}{\sum_{m=1}^{M}\sum_{t=1}^{T} P(\boldsymbol{O}, i_t = m \mid \lambda')} \\ &= \frac{1}{T}\sum_{t=1}^{T} P(i_t = m \mid \boldsymbol{O}, \lambda') \\ &= \frac{1}{T}\sum_{t=1}^{T} \gamma'_t(m), \end{aligned} \tag{14}$$

where $\gamma_t(m)$ is the posterior probability of being in the m-th mixture component at time t, that is

$$\begin{aligned} \gamma_t(m) &= P(i_t = m \mid \boldsymbol{O}, \lambda) \\ &= \frac{c_m \prod_{s=1}^{S} p_{ms}(\boldsymbol{o}_{ts})}{b(\boldsymbol{o}_t)}. \end{aligned} \tag{15}$$

Similarly, the second term is maximized as

$$w_{msg} = \frac{\sum_{t \in T(\boldsymbol{O}, s, g)} \xi'_{ts}(m, g)}{\sum_{l=1}^{G_s}\sum_{t \in T(\boldsymbol{O}, s, l)} \xi'_{ts}(m, l)}, \tag{16}$$

where $\xi_{ts}(m, g)$ is the posterior probability of being in the g-th space of stream s in the m-th mixture componentat time t:

$$\begin{aligned}\xi_{ts}(m,g) &= P(i_t = m, l_{ts} = g \mid \boldsymbol{O}, \lambda) \\ &= P(i_t = m \mid \boldsymbol{O}, \lambda) P(l_{ts} = g \mid i_t = m, \boldsymbol{O}, \lambda) \\ &= \gamma_t(m) \frac{w_{msg} \mathcal{N}_{msg}^{D_{sg}}(\boldsymbol{x}_{ts})}{p_{ms}(\boldsymbol{o}_{ts})}. \end{aligned} \tag{17}$$

The third term is maximized by solving following equations:

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{\mu}_{msg}} \sum_{t \in T(\boldsymbol{O},s,g)} P(\boldsymbol{O}, i_t = m, l_{ts} = g \mid \lambda') \\ &\quad \cdot \log \mathcal{N}_{msg}^{D_{sg}}(\boldsymbol{x}_{ts}) = \boldsymbol{0}, \end{aligned} \tag{18}$$

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{\Sigma}_{msg}^{-1}} \sum_{t \in T(\boldsymbol{O},s,g)} P(\boldsymbol{O}, i_t = m, l_{ts} = g \mid \lambda') \\ &\quad \cdot \log \mathcal{N}_{msg}^{D_{sg}}(\boldsymbol{x}_{ts}) = \boldsymbol{0}, \end{aligned} \tag{19}$$

resulting in

$$\boldsymbol{\mu}_{msg} = \frac{\displaystyle\sum_{t \in T(\boldsymbol{O},s,g)} \xi'_{ts}(m,g) \boldsymbol{x}_{ts}}{\displaystyle\sum_{l=1}^{G_s} \sum_{t \in T(\boldsymbol{O},s,l)} \xi'_{ts}(m,l)}, \tag{20}$$

$$\begin{aligned}&\boldsymbol{\Sigma}_{msg} \\ &= \frac{\displaystyle\sum_{t \in T(\boldsymbol{O},s,g)} \xi'_{ts}(m,g)(\boldsymbol{x}_{ts} - \boldsymbol{\mu}_{msg})(\boldsymbol{x}_{ts} - \boldsymbol{\mu}_{msg})^\top}{\displaystyle\sum_{l=1}^{G_s} \sum_{t \in T(\boldsymbol{O},s,l)} \xi'_{ts}(m,l)}. \end{aligned} \tag{21}$$

The re-estimation is repeated iteratively using λ in place of λ' and the final result is an ML estimation of the MSD-GMM.

4. MCE Training for MSD-GMMs

The generalized probabilistic descent (GPD) method optimizes the parameters of a classifier in a pattern recognizer using a gradient technique [12], [13]. GPD has also been applied to speaker recognition tasks and several works have demonstrated the utility of GPD in speaker recognition [14]–[17].

In this paper, we apply the GPD method to MSD-GMM-based speaker identification. In this case, the adjustable parameter set Λ includes the entire parameters of N MSD-GMM speaker models

$$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}. \tag{22}$$

GPD minimizes the recognition error probability by iteratively minimizing an objective function called "*empirical loss*"which is a good approximation of the recognizer's error rate over the training data.

4.1 Definition of Empirical Loss

The *empirical loss* can be defined based on a smooth embedding of three functions: *discriminant function*, *misclassification measure* and *loss function*.

We define the *discriminant function* for speaker class C_n upon observing feature vector sequence $\boldsymbol{O}$ as the average log-likelihood of λ_n:

$$\begin{aligned} g_n(\boldsymbol{O}; \Lambda) &= \frac{1}{T} \log P(\boldsymbol{O} \mid \lambda_n) \\ &= \frac{1}{T} \sum_{t=1}^{T} \log b(\boldsymbol{o}_t \mid \lambda_n). \end{aligned} \tag{23}$$

The *misclassification measure* for C_n is defined by using the discriminant functions of speaker n and the dominant competing speaker j:

$$d_n(\boldsymbol{O}; \Lambda) = -g_n(\boldsymbol{O}; \Lambda) + g_j(\boldsymbol{O}; \Lambda), \tag{24}$$

$$j = \arg \max_{i \neq n} g_i(\boldsymbol{O}; \Lambda), \tag{25}$$

where $d_n > 0$ implies misclassification and $d_k \leq 0$ means correct classification.

Then, the *loss function* is defined as a differentiable sigmoid function approximating the 0-1 step loss function:

$$\ell_n(\boldsymbol{O}; \Lambda) = \frac{1}{1 + \exp\{-\beta d_n(\boldsymbol{O}; \Lambda)\}} \tag{26}$$

The *empirical loss* is defined as the average of loss functions over the training data set $\mathcal{O} = \{\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_H\}$:

$$\mathcal{L}(\mathcal{O}; \Lambda) = \frac{1}{H} \sum_{h=1}^{H} \sum_{n=1}^{N} \ell_n(\boldsymbol{O}_h; \Lambda) 1(\boldsymbol{O}_h \in C_n), \tag{27}$$

where $1(\mathcal{X})$ is the indicator function for a logical variable $\mathcal{X}$ defined as

$$1(\mathcal{X}) = \begin{cases} 1, & \text{if } \mathcal{X} \text{ is true} \\ 0, & \text{otherwise} \end{cases}. \tag{28}$$

4.2 Minimization of Empirical Loss

The parameter set Λ is sequentially adjusted every time a training sample $\boldsymbol{O}_h$ from speaker class C_k is given according to

$$\Lambda^{(r+1)} = \Lambda^{(r)} - \varepsilon^{(r)} \boldsymbol{U}^{(r)} \nabla_\Lambda \ell_k(\boldsymbol{O}_h; \Lambda)\big|_{\Lambda = \Lambda^{(r)}}, \tag{29}$$

where $\varepsilon^{(r)}$ is a monotonically decreasing learning step size and $\boldsymbol{U}^{(r)}$ is a positive definite learning matrix at the r-th iteration.

During the parameter adaptation, the constraints of MSD-GMM parameters, such as $c_m > 0$ and $w_{msg} > 0$, should be satisfied, and means are normalized with

variances. The following parameter transformations allow us to maintain these constraints:

$$\bar{c}_m = \log c_m, \tag{30}$$

$$\bar{w}_{msg} = \log w_{msg}, \tag{31}$$

$$\bar{\mu}_{msgd} = \frac{\mu_{msgd}}{\sigma_{msgd}}, \tag{32}$$

$$\bar{\sigma}^2_{msgd} = \log \sigma^2_{msgd}, \tag{33}$$

where μ_{msgd} is the d-th element of $\boldsymbol{\mu}_{msg}$ and σ^2_{msgd} is the d-th diagonal element of (diagonal) matrix $\boldsymbol{\Sigma}_{msg}$. Using the new parameter set $\bar{\Lambda}$, (29) is rewritten as

$$\bar{\Lambda}^{(r+1)} = \bar{\Lambda}^{(r)} - \varepsilon^{(r)} \boldsymbol{U}^{(r)} \nabla_{\bar{\Lambda}} \ell_k(\boldsymbol{O}_h; \Lambda)\big|_{\Lambda=\Lambda^{(r)}}. \tag{34}$$

The parameter sequence produced by (34) converges (with probability one) to the minimum point of (27) for large H [12].

According to the chain rule, the gradient $\nabla_{\bar{\Lambda}} \ell_k$ (the subscript h is omitted and $\ell_k(\boldsymbol{O}; \Lambda)$ is shortened to ℓ_k to simplify the notation) in (34) can be rewritten as

$$\nabla_{\bar{\Lambda}} \ell_k = \frac{\partial \ell_k}{\partial d_k} \sum_{n=1}^{N} \frac{\partial d_k}{\partial g_n} \sum_{t=1}^{T} \frac{\partial g_n}{\partial b(\boldsymbol{o}_t \mid \lambda_n)} \nabla_{\bar{\Lambda}} b(\boldsymbol{o}_t \mid \lambda_n), \tag{35}$$

where

$$\frac{\partial \ell_k}{\partial d_k} = \beta\, \ell_k \left(1 - \ell_k\right), \tag{36}$$

$$\frac{\partial d_k}{\partial g_n} = \begin{cases} -1, & n = k \\ 1, & n = \arg\max\limits_{i \neq k} g_i \\ 0, & \text{otherwise} \end{cases}, \tag{37}$$

$$\frac{\partial g_n}{\partial b(\boldsymbol{o}_t \mid \lambda_n)} = \frac{1}{T b(\boldsymbol{o}_t \mid \lambda_n)}. \tag{38}$$

Dropping the subscripts t and n for simplicity of notation, each component of $\nabla_{\bar{\Lambda}} b(\boldsymbol{o} \mid \lambda)$ can be derived as follows:

$$\frac{\partial b(\boldsymbol{o} \mid \lambda)}{\partial \bar{c}_m} = c_m \prod_{s=1}^{S} p_{ms}(\boldsymbol{o}_s). \tag{39}$$

For $g \in X_s$,

$$\frac{\partial b(\boldsymbol{o} \mid \lambda)}{\partial \bar{w}_{msg}} = \zeta_{msg}(\boldsymbol{o}), \tag{40}$$

$$\frac{\partial b(\boldsymbol{o} \mid \lambda)}{\partial \bar{\mu}_{msgd}} = \frac{x_{sd} - \mu_{msgd}}{\sigma_{msgd}} \zeta_{msg}(\boldsymbol{o}), \tag{41}$$

$$\frac{\partial b(\boldsymbol{o} \mid \lambda)}{\partial \bar{\sigma}^2_{msgd}} = \frac{1}{2} \left\{ \left(\frac{x_{sd} - \mu_{msgd}}{\sigma_{msgd}} \right)^2 - 1 \right\} \zeta_{msg}(\boldsymbol{o}), \tag{42}$$

where x_{sd} is the d-th element of vector $\boldsymbol{x}_s$ and

$$\zeta_{msg}(\boldsymbol{o}) = c_m w_{msg} \mathcal{N}^{D_{sg}}_{msg}(\boldsymbol{x}_s) \prod_{h \neq s} p_{mh}(\boldsymbol{o}_h). \tag{43}$$

For $g \notin X_s$,

$$\frac{\partial b(\boldsymbol{o} \mid \lambda)}{\partial \bar{w}_{msg}} = \frac{\partial b(\boldsymbol{o} \mid \lambda)}{\partial \bar{\mu}_{msgd}} = \frac{\partial b(\boldsymbol{o} \mid \lambda)}{\partial \bar{\sigma}^2_{msgd}} = 0. \tag{44}$$

After the adjustment of $\bar{\Lambda}$ is completed, $\bar{\Lambda}$ is transformed back to Λ as follows:

$$c_m = \frac{\exp \bar{c}_m}{\sum\limits_{i=1}^{M} \exp \bar{c}_i}, \tag{45}$$

$$w_{msg} = \frac{\exp \bar{w}_{msg}}{\sum\limits_{l=1}^{G_s} \exp \bar{w}_{msl}}, \tag{46}$$

$$\mu_{msgd} = \sigma_{msgd} \bar{\mu}_{msgd}, \tag{47}$$

$$\sigma^2_{msgd} = \exp \bar{\sigma}^2_{msgd}. \tag{48}$$

5. Experimental Evaluation

5.1 Databases

Text-independent speaker identification experiments were carried out using the ATR Japanese speech database [18] and the NTT VR database [19].

We used word data spoken by 80 speakers (40 males and 40 females) in "c-set" of the ATR 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 NTT database consists of sentence data uttered at three speeds (normal, fast, and slow) by 35 Japanese speakers (22 males and 13 females) on five sessions over ten months (Aug., Sept., Dec. 1990, Mar., June 1991), among which, the normal-speed data set was used. In each session, 15 sentences were recorded for each speaker. Ten sentences are common to all speakers and all sessions (A-set), and five sentences are different for each speaker and each session (B-set). The duration of each sentence is approximately four second. We used 15 sentences (A-set + B-set from the first session) per speaker for training, and 20 sentences (B-set from the other four sessions) per speaker for testing. The number of tests was 700 in total.

5.1.1 Speech Analysis and Training

The speech data were down-sampled 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 [20]. The 12 static parameters excluding the zero-th coefficient were used as a spectral feature. Fundamental frequencies (F_0) were estimated at a 10-ms frame rate using the RAPT method [21] with a 7.5-ms correlation

window, and $\log F_0$ for the voiced frames and discrete symbols for unvoiced frames were used as a pitch feature.

Speakers were modeled by GMMs or multi-stream MSD-GMMs with diagonal covariance matrices. The baseline GMM and MSD-GMM parameters for MCE training were initialized with an LBG codebook and ML-trained. Identity matrices were used for the learning matrices $\boldsymbol{U}^{(r)}$ and the learning step size $\varepsilon^{(r)}$ was initialized as $\varepsilon^{(0)} = 0.2$. The slope of the sigmoid function β for each model set Λ was estimated before the training using the variance of the misclassification measures for all the training samples v, according to

$$\beta = \frac{4}{\sqrt{2\pi v}}. \tag{49}$$

The resulting values of β were about $1.0 - 2.0$. The GPD training was iterated over 20 epochs, with the order of the given training samples being shuffled at the beginning of each epoch.

5.2 Experimental Results

5.2.1 Comparison with Conventional Systems

We compared the MSD-GMM speaker identification system with three kinds of conventional systems. Figure 4 shows speaker identification error rates for the ATR database when using 32 and 64 component ML-trained speaker models with 95% confidence intervals (CIs). In the figure, "GMM" denotes a conventional GMM speaker model using a spectral feature alone, "S+P-GMM" represents a speaker model consisting of two GMMs for spectra and pitch, "V+UV-GMM" is the speaker model consisting of two GMMs for voiced (V) and unvoiced (UV) parts [8], with a linear combination parameter $\alpha = 0.5$ (α is the weight for the likelihood of the UV-GMM), and "MSD-GMM" denotes the proposed model based on the multi-stream MSD-GMM. For the "V+UV-GMM" system, the optimum shares of the mixture components for the V-GMM and the UV-GMM were tuned as in Fig. 5, where the best results were obtained with the ratio V : UV = 3 : 1.

As shown in Fig. 4, the additional use of pitch information significantly improved the system performance, and S+P-GMM, V+UV-GMM and MSD-GMM 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. The MSD-GMM system achieved 16% and 18% error reductions over the GMM system for the 32 and 64 mixture models, respectively.

The differences between the above three models can be summarized as follows. S+P-GMM ignores the synchronousness of spectra and pitch, and also ignores

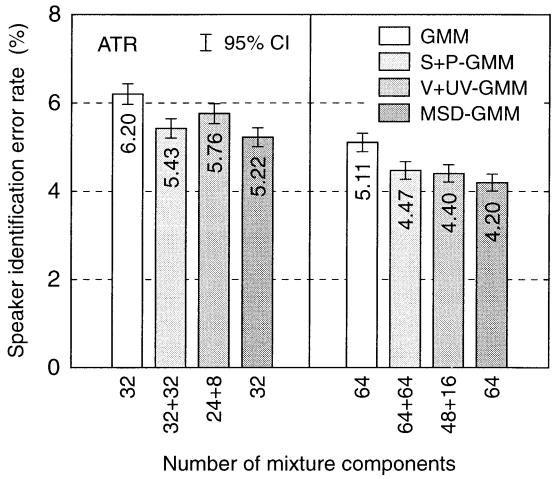


Fig. 4 Comparison of MSD-GMM speaker models with conventional GMM speaker models (ATR database).

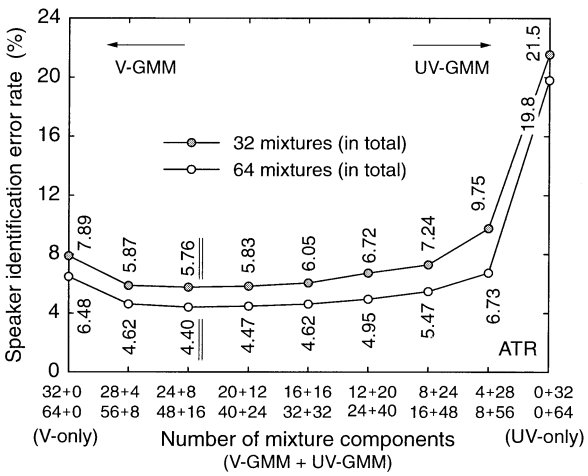


Fig. 5 Speaker identification error rates for the V+UV-GMM system with the ratio V : UV (ATR database).

the weights for voiced and unvoiced parts. V+UV-GMM maintains the synchronousness of spectra and pitch, and can adjust the weights for voiced and unvoiced parts by tuning the numbers of mixture components for V-GMM and UV-GMM or the linear combination parameter (α). MSD-GMM also maintains the synchronousness of spectra and pitch, and can adjust the weights for voiced and unvoiced parts more flexibly than V+UV-GMM by estimating space weights in each stream of each mixture. It is also noted that the MSD-GMM system requires no combination parameter (such as α) which has to be chosen or tuned heuristically.

To evaluate the robustness of the MSD-GMM speaker model for inter-session variability, we also conducted speaker identification experiments using the NTT database with session-to-session variation. The results are shown in Fig. 6. Similar results were obtained for the NTT database and the MSD-GMM system achieved error reductions of 38% and 51% over the

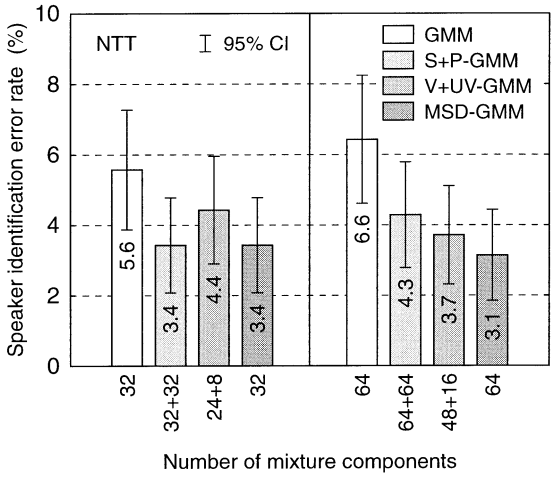


Fig. 6 Comparison of MSD-GMM speaker models with conventional GMM speaker models (NTT database).

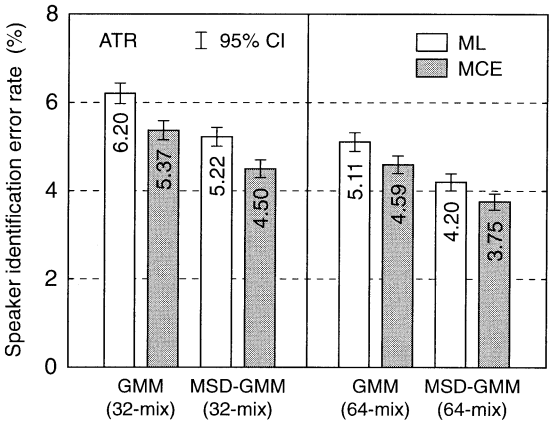


Fig. 7 Comparison of ML- and MCE-based GMM and MSD-GMM systems (ATR database).

GMM system for the case of 32 and 64 mixtures, respectively.

5.2.2 Comparison of Training Methods

Speaker models were trained based on two different methods: ML estimation and MCE training described in Sects. 3 and 4, respectively. The results for the ATR database are shown in Fig. 7, which compares ML- and MCE-based GMM and MSD-GMM systems. The ML-based systems are identical to those in Fig. 4. In both cases of GMM and MSD-GMM, the system performance is significantly improved by MCE training. The MCE-based MSD-GMM system gave the lowest error rates in both model sizes, and achieved 14% and 11% error reductions over the ML-based MSD-GMM system, yielding 28% and 26% error reductions over the ML-based GMM system for the 32 and 64 mixture models, respectively. There is no overlap of CIs between the MCE-based MSD-GMM system and the other three systems.

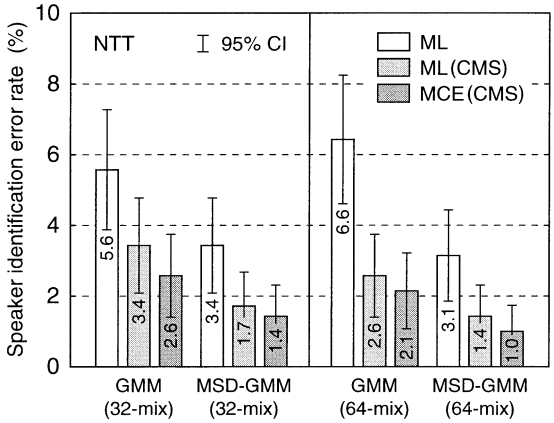


Fig. 8 Comparison of ML- and MCE-based GMM and MSD-GMM systems (NTT database).

The robustness of MCE training is also evaluated for the multi-session NTT database. The results are shown in Fig. 8. MCE training is applied after normalizing the session-dependent utterance variation using cepstrum mean subtraction (CMS) method, which is a well-known technique for canceling the effect of channels and utterance variation in speaker recognition [22], [23]. System performance was significantly improved with CMS effectively canceling the inter-session variability. MCE training further reduced the identification errors and an error rate of 1.0% was obtained for the 64-mixture MSD-GMM.

6. Conclusion

This paper has introduced a new technique for modeling speakers based on MSD-GMM for text-independent speaker identification. MSD-GMM can model continuous pitch values of voiced frames and discrete symbols representing "unvoiced frame" in a unified framework. Spectral and pitch features can be jointly modeled by a multi-stream MSD-GMM. We derived the ML estimation formulae and the MCE training procedure for the MSD-GMM parameters and evaluated the MSD-GMM speaker models for text-independent speaker identification tasks. The experimental results demonstrated that the MSD-GMM can efficiently model each speaker and the identification accuracy was improved by the use of pitch information along with spectral information. The error rates were further reduced with the MCE training of the MSD-GMM parameters. Furthermore, the results for the multi-session database proved the robustness of the MSD-GMM speaker model against the inter-session variability.

Alternately, spectral and pitch features can be modeled by a single-stream MSD-GMM with two spaces; the first space corresponds to the voiced frames whose observation vectors are the concatenations of spectral and pitch vectors, and the second space is

for the unvoiced frames whose observation vectors are spectral vectors alone. We can also introduce stream weights into the MSD-GMM, and the stream weights as well as MSD-GMM parameters can be estimated based on MCE training. Comparison to such speaker models and the application of this framework to a speaker verification system will be the subjects for future works.

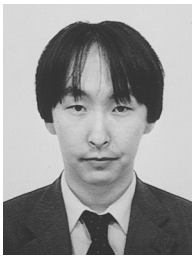
Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions on this article. The authors would also like to thank Professor Sadaoki Furui of Tokyo Institute of Technology, and Dr. Tomoko Matsui of ATR Spoken Language Translation Laboratories, for providing the NTT database. A part of this work was supported by Research Fellowships from the Japan Society for the Promotion of Science for Young Scientists, No.199808177.

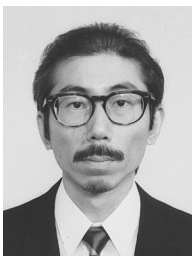
References

[1] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech & Audio Process., vol.3, no.1, pp.72–83, Jan. 1995.
[2] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, and B.-H. Juang, "A vector quantization approach to speaker recognition," Proc. ICASSP'85, vol.1, pp.387–390, March 1985.
[3] B.S. Atal, "Automatic speaker recognition based on pitch contours," J. Acoust. Soc. Am., vol.52, no.6, pp.1687–1697, 1972.
[4] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques," Speech Commun., vol.5, no.2, pp.183–197, 1986.
[5] M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," Proc. ICSLP'96, vol.3, pp.1800–1804, Oct. 1996.
[6] M.K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," Proc. EUROSPEECH'97, vol.3, pp.1391–1394, Sept. 1997.
[7] T. Matsui and S. Furui, "Text-independent speaker recognition using vocal tract and pitch information," Proc. ICSLP'90, vol.1, pp.137–140, Nov. 1990.
[8] K.P. Markov and S. Nakagawa, "Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition," J. Acoust. Soc. Jpn. (E), vol.20, no.4, pp.281–291, July 1999.
[9] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," Proc. ICASSP'99, vol.1, pp.229–232, May 1999.
[10] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," IEICE Trans., vol.J83-D-II, no.7, pp.1579–1589, July 2000.
[11] C. Miyajima, Y. Hattori, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker identification using Gaussian mixture models based on multi-space probability distribution," Proc. ICASSP 2001, vol.1, May 2001.
[12] 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.
[13] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," Proc. IEEE, vol.86, no.11, pp.2345–2373, Nov. 1998.
[14] C.-S. Liu, C.-H. Lee, W. Chou, 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.
[15] C.M. del Álamo, F.J.C. Gil, C. de la T. Munilla, and L.H. Gómez, "Discriminative training of GMM for speaker identification," Proc. ICASSP'96, vol.1, pp.89–92, May 1996.
[16] A. Rosenberg, O. Siohan, and S. Parthasarathy, "Speaker verification using minimum verification error training," Proc. ICASSP'98, vol.1, pp.105–108, May 1998.
[17] O. Siohan, A.E. Rosenberg, and S. Parthasarathy, "Speaker identification using minimum classification error training," Proc. ICASSP'98, vol.1, pp.109–112, May 1998.
[18] H. Kuwabara, Y. Sagisaka, K. Takeda, and M. Abe, "Construction of ATR Japanese speech database as a research tool (Appendix II)," ATR Interpreting Telephony Research Laboratories Technical Report, TR-I-0086, June 1989.
[19] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," Proc. ICASSP'92, vol.2, pp.157–160, March 1992.
[20] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. ICASSP'92, vol.1, pp.137–140, March 1992.
[21] D. Talkin, "A robust algorithm for pitch tracking," in Speech Coding and Synthesis, eds. W.B. Kleijn and K.K. Paliwal, pp.495–518, Elsevier Science, 1995.
[22] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., vol.55, no.6, pp.1304–1312, 1974.
[23] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust. Speech & Signal Process., vol.29, no.2, pp.254–272, April 1981.

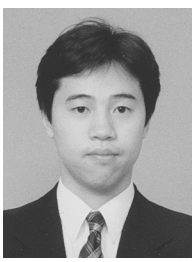
Chiyomi Miyajima received the B.E. degree in computer science and M.E. and Dr.Eng. degrees in electrical and computer engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1996, 1998 and 2001, respectively. Currently she is a Research Associate of the Department of Computer Science, Nagoya Institute of Technology. Her research interests include automatic speaker recognition and multimodal speech recognition. She received Kiyoshi Awaya Award in 2000 from the Acoustical Society of Japan. She is a member of ASJ.



Yosuke Hattori received the B.E. degree in computer science and M.E. degree in electrical and computer engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1998 and 2000, respectively. He is currently with DENSO Corporation.



Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1984, and M.E. and Dr.Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1986 and 1989, respectively. From 1989 to 1996, he was a Research Associate at Department of Electronic and Electric Engineering, Tokyo Institute of Technology. Since 1996 he has been with the Department of Computer Science, Nagoya Institute of Technology as an Associate Professor. His research interests include speech spectral estimation, speech coding, speech synthesis and recognition, and adaptive signal processing. He is a member of IEEE, ASJ, IPSJ and JSAI.



Takashi Masuko received the B.E. degree in computer science, and M.E. degree in intelligence science from Tokyo Institute of Technology, Tokyo, Japan, in 1993 and 1995, respectively. In 1995, he joined the Precision and Intelligence Laboratory, Tokyo Institute of Technology as a Research Associate. He is currently a Research Associate of the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. His research interests include speech synthesis, speech recognition, speech coding, and multimodal interface. He is a member of IEEE, ISCA and ASJ.

Takao Kobayashi received the B.E. degree in electrical engineering, the M.E. and Dr.Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1977, 1979, and 1982, respectively. In 1982, he joined the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology as a Research Associate. He became an Associate Professor at the same Laboratory in 1989. He is currently a Professor of the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. His research interests include speech analysis and synthesis, speech coding, speech recognition, and multimodal interface. He is a member of IEEE, ISCA and ASJ.

Tadashi Kitamura received the B.E. degree in electrical and electronic engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1973, and M.E. and Dr.Eng. degrees from Tokyo Institute of Technology, Tokyo, Japan, in 1975 and 1978, respectively. He is currently a Professor of Department of Computer Science, Nagoya Institute of Technology. His research interests include speech information processing, multi-modal information processing and biometric person authentication. He is a member of IEEE, ISCA, EURASIP, ASJ and IPSJ.