IEICE TRANS. INF. & SYST., VOL.E94–D, NO.1 JANUARY 2011

LETTER

Artificial Cohort Generation Based on Statistics of Real Cohorts for GMM-Based Speaker Verification

Yuuji MUKAI[†], *Nonmember*, **Hideki NODA**[†a)], *and* **Takashi OSANAI**[††], *Members*

SUMMARY This paper discusses speaker verification (SV) using Gaussian mixture models (GMMs), where only utterances of enrolled speakers are required. Such an SV system can be realized using artificially generated cohorts instead of real cohorts from speaker databases. This paper presents a rational approach to set GMM parameters for artificial cohorts based on statistics of GMM parameters for real cohorts. Equal error rates for the proposed method are about 10% less than those for the previous method, where GMM parameters for artificial cohorts were set in an ad hoc manner.

key words: *speaker verification, Gaussian mixture model, artificial cohort, score normalization, background model*

1. Introduction

Speaker verification (SV) is the task of determining whether the claimed identity of a speaker is correct. Statistical approaches using Gaussian mixture models (GMMs) are commonly used for text-independent SV [1]. An important issue in the statistical approaches is that of score normalization, and several normalization methods have already been proposed. Popular normalization methods include normalization using universal background model [1], cohort normalization method [2], and T-norm [3].

All of the above score normalization methods need speaker databases. However, preparation of a database is a burden particularly in a small-scale SV system such as one for home security, where the number of enrolled speakers is very small. In such an SV system, it is desirable to perform SV using only utterances of enrolled speakers. This issue was already addressed in text-dependent SV using hidden Markov model [4] and text-independent SV using GMM [5], [6].

In [5], a background model was estimated using the training data for a claimed speaker, i.e., the same data was used to build the claimed speaker model and its background model. The difference between the two models is the number of Gaussian mixtures, where the smaller number was used for the background model. In [6], artificial cohorts were used instead of real cohorts from speaker databases. Considering that GMMs for cohorts for a claimed speaker are relatively close to that for the claimed speaker, GMMs for cohorts were generated by changing model parameters of the GMM for the claimed speaker. SV performance by this approach was better than that by [5]. However the GMMs for cohorts were generated in an ad hoc manner, i.e., variation ranges of GMM parameters for cohorts were set merely experimentally. Alternatively, this paper proposes a rational approach to set GMM parameters for artificial cohorts based on statistics of GMM parameters for real cohorts.

The use of artificial cohorts can be most effective in case that the universal background model or real cohorts from a database cannot be available in an SV system because of the recording condition (including transmission channel) for the SV system being different from that for the database. Assuming that a different recording condition changes feature vectors only by adding a speaker independent vector, the difference between GMM for a claimed speaker and that for a cohort is considered to be independent of the recording condition. Then, once necessary parameters to generate artificial cohorts (see Sect. 4) have been estimated based on statistics of GMM parameters for real cohorts from a database, we can expect that those parameters can be utilized in any SV system.

2. Score Normalization Methods for Speaker Verification

Let $\mathbf{Y} = \{\mathbf{y}_t; t = 1, \cdots, T\}$ denote a sequence of feature vectors obtained from input speech, and let $p_s(\mathbf{y}_t)$ and $p_o(\mathbf{y}_t)$ be probability density functions (pdfs) of $\mathbf{y}_t$ for a claimed speaker (true speaker) and all other possible speakers (impostors), respectively. Here both pdfs are modeled by GMMs. Assuming that $\mathbf{y}_t$s are mutually independent and then $p_s(\mathbf{Y}) = \prod_{t=1}^{T} p_s(\mathbf{y}_t)$ and $p_o(\mathbf{Y}) = \prod_{t=1}^{T} p_o(\mathbf{y}_t)$, log-likelihood ratio $S(\mathbf{Y})$ is given as

$$S(\mathbf{Y}) = \log \frac{p_s(\mathbf{Y})}{p_o(\mathbf{Y})} \tag{1}$$

$$= \sum_{t=1}^{T} \log \frac{p_s(\mathbf{y}_t)}{p_o(\mathbf{y}_t)}. \tag{2}$$

In fact, instead of $p_s(\mathbf{Y})$ and $p_o(\mathbf{Y})$, the normalized likelihood by the length T of the vector sequence $\mathbf{Y}$ (the number of frames), i.e., $p_s(\mathbf{Y})^{1/T}$ and $p_o(\mathbf{Y})^{1/T}$ are usually used. In that case, the log-likelihood ratio $S(\mathbf{Y})$ is given as

$$S(\mathbf{Y}) = \frac{1}{T} \log \frac{p_s(\mathbf{Y})}{p_o(\mathbf{Y})} \tag{3}$$

Manuscript received June 22, 2010.
Manuscript revised August 30, 2010.

[†]The authors are with the Department of Systems Design and Informatics, Kyushu Institute of Technology, Iizuka-shi, 820–8502 Japan.
[††]The author is with National Research Institute of Police Science, Kashiwa-shi, 277–0882 Japan.
a) E-mail: noda@mip.ces.kyutech.ac.jp

DOI: 10.1587/transinf.E94.D.162

$$= \frac{1}{T}\sum_{t=1}^{T}\log\frac{p_s(\mathbf{y}_t)}{p_o(\mathbf{y}_t)}. \quad (4)$$

Using S(**Y**), the decision on the hypothesis that **Y** is from the claimed speaker is made as follows.

$$S(\mathbf{Y}) \geq \theta_{th}, \text{ accept the hypothesis} \quad (5)$$

$$S(\mathbf{Y}) < \theta_{th}, \text{ reject the hypothesis,} \quad (6)$$

where θ_{th} is a decision threshold.

Taking the aforementioned general procedure into account, two popular score normalization methods are described as follows.

(1) Use of universal background model [1]
Speech samples from a large number of speakers are used to train a single GMM for $p_o(\mathbf{y}_t)$, which is called a universal background model or a world model.

(2) Cohort normalization [2]
Cohort normalization uses a set of other speakers called cohorts whose pdfs are close to that for a claimed speaker. Cohorts for the claimed speaker are selected from speaker databases. Given the selected cohorts $c_i, i = 1, \cdots, N$ and their pdfs $p_{c_i}(\mathbf{Y})$, the following $p_o(\mathbf{Y})$,

$$p_o(\mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N} p_{c_i}(\mathbf{Y}) \quad (7)$$

is used instead of the universal background model.

3. Previous Artificial Cohort Model

In this section, after a brief introduction of GMM, we review the previous method to generate GMMs for artificial cohorts [6].

A mixture of K Gaussian distributions is described as

$$p(\mathbf{y}_t) = \sum_{k=1}^{K} a_k g_k(\mathbf{y}_t; \mathbf{m}_k, \mathbf{\Sigma}_k), \quad \sum_{k=1}^{K} a_k = 1, \quad (8)$$

$$g_k(\mathbf{y}_t; \mathbf{m}_k, \mathbf{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_k|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(\mathbf{y}_t - \mathbf{m}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{y}_t - \mathbf{m}_k)\right\}, \quad (9)$$

where $\mathbf{y}_t$ is a D dimensional feature vector at t-th frame and a_k is the mixing coefficient of the k-th Gaussian distribution $g_k(\mathbf{y}_t; \mathbf{m}_k, \mathbf{\Sigma}_k)$ with mean vector $\mathbf{m}_k$ and covariance matrix $\mathbf{\Sigma}_k$. The model parameters, $a_k, \mathbf{m}_k, \mathbf{\Sigma}_k, k = 1, \cdots, K$ are iteratively estimated by the EM method [7]. Explicit procedures of the EM method are found in [8], [9]. The initial values to start the iterative procedure are obtained by clustering training samples using the VQ method [10].

Given model parameters, $a_k, \mathbf{m}_k, \mathbf{\Sigma}_k, k = 1, \cdots, K$ of the GMM $p_s(\mathbf{y}_t)$ for a claimed speaker, those parameters $a_k^{c_i}, \mathbf{m}_k^{c_i}, \mathbf{\Sigma}_k^{c_i}$ for its artificial cohorts $c_i, i = 1, \cdots, N$ were set as

$$a_k^{c_i} = a_k, \quad (10)$$

$$\mathbf{m}_k^{c_i} = \mathbf{m}_k + \alpha\mathbf{r}_k^{c_i}, \quad (11)$$

$$\mathbf{\Sigma}_k^{c_i} = \beta\mathbf{\Sigma}_k, \quad (12)$$

where α and β are parameters which should be set experimentally, and $\mathbf{r}_k^{c_i}$ is a random vector whose components $r_{k,d}^{c_i}, d = 1, \cdots, D$ are uniformly distributed in the interval $-1 \leq r_{k,d}^{c_i} \leq 1$. The parameter α controls variations of $\mathbf{m}_k^{c_i}$ for cohorts from $\mathbf{m}_k$, and the parameter $\beta > 1$ increases variances for cohorts from those for the claimed speaker. $\alpha = 0.2$ and $\beta = 2$ were used in [6].

4. Artificial Cohort Model Based on Statistics of Real Cohorts

We take a rational approach where GMM parameters for artificial cohorts are set based on statistics of GMM parameters for real cohorts. In order to realize it, we investigate the differences of the GMM parameters between claimed speakers and their real cohorts. To investigate those differences, each of mixed Gaussians for a claimed speaker needs to have a corresponding Gaussian among mixed ones for each cohort. We find out such a correspondence relation using the Kullback-Leibler (KL) divergence. The KL divergence D_{KL} for two Gaussians, $g_1(\mathbf{y}; \mathbf{m}_1, \mathbf{\Sigma}_1)$ and $g_2(\mathbf{y}; \mathbf{m}_2, \mathbf{\Sigma}_2)$ is given as [11]

$$D_{KL} = \int_{-\infty}^{\infty} g_1(\mathbf{y}; \mathbf{m}_1, \mathbf{\Sigma}_1)\ln\frac{g_1(\mathbf{y}; \mathbf{m}_1, \mathbf{\Sigma}_1)}{g_2(\mathbf{y}; \mathbf{m}_2, \mathbf{\Sigma}_2)}d\mathbf{y} \quad (13)$$

$$= \frac{1}{2}tr(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1 - \mathbf{I}) + \frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{\Sigma}_2^{-1}(\mathbf{m}_1 - \mathbf{m}_2) + \frac{1}{2}\ln\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}. \quad (14)$$

Taking one Gaussian from the mixtures for a claimed speaker as the above $g_1(\mathbf{y}; \mathbf{m}_1, \mathbf{\Sigma}_1)$ and taking one Gaussian among the mixtures for a cohort as the above $g_2(\mathbf{y}; \mathbf{m}_2, \mathbf{\Sigma}_2)$, the one among the mixtures for a cohort which gives the smallest KL divergence is decided as the corresponding one for that Gaussian for a claimed speaker.

Let $m_{k,d}$ and $m_{k',d}^{c_i}$ be the d-th component of the mean vector $\mathbf{m}_k$ of k-th Gaussian for a claimed speaker and that of $\mathbf{m}_{k'}^{c_i}$ of the corresponding k'-th Gaussian for its real cohorts $c_i, i = 1, \cdots, N$, respectively, and $\sigma_{k,d}$ and $\sigma_{k',d}^{c_i}$ be the d-th diagonal component of the covariance matrix $\mathbf{\Sigma}_k$ for a claimed speaker and that of $\mathbf{\Sigma}_{k'}^{c_i}$ for its real cohorts, respectively, assuming that the covariance matrices are diagonal. The data of $m_{k',d}^{c_i} - m_{k,d}$ and $1/(\sigma_{k',d}^{c_i}/\sigma_{k,d}) = \sigma_{k,d}/\sigma_{k',d}^{c_i}$ for all speakers are collected, where the number of data is the number of speakers × the number of cohorts per each speaker × the number of mixed Gaussians. As an example, the distribution of $m_{k',d}^{c_i} - m_{k,d}$ and that of $\sigma_{k,d}/\sigma_{k',d}^{c_i}$ both for fifth component ($d = 5$) are shown in Fig. 1 and Fig. 2, respectively. According to the experimental conditions described in Sect. 5, where the number of speakers is 100, the number of cohorts per each speaker is 50, and the number of mixed Gaussians is 16, the number of data used for these figures is

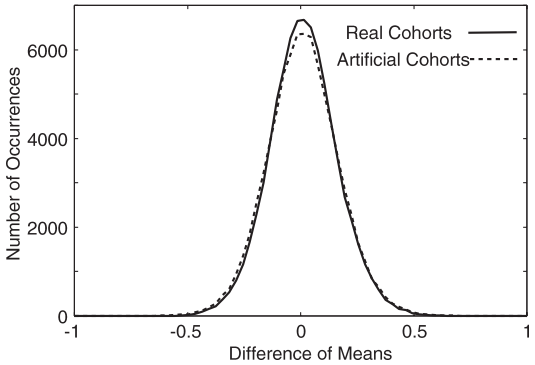

Fig. 1 Distribution of difference of means $m_{k',d}^{c_i} - m_{k,d}$ for $d = 5$ for real and artificial cohorts.

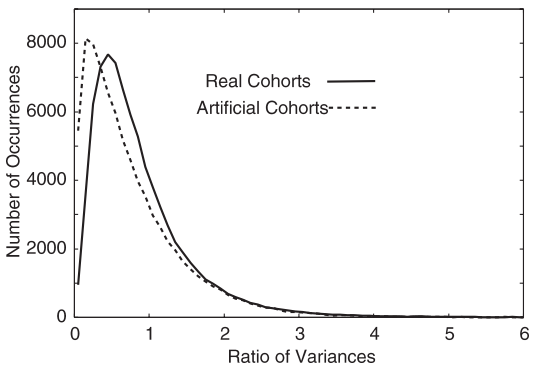

Fig. 2 Distribution of ratio of two variances $\sigma_{k,d}/\sigma_{k',d}^{c_i}$ for $d = 5$ for real and artificial cohorts.

Table 1 Estimated parameters: standard deviation parameter λ for Gaussian distribution and parameters θ_1 and θ_2 for gamma distribution.

	d (component number)					
	1	3	5	7	9	11
λ	0.300	0.210	0.150	0.124	0.098	0.088
θ_1	1.318	1.586	1.716	1.576	2.156	1.774
θ_2	0.691	0.555	0.509	0.567	0.409	0.495

$100\times50\times16 = 80,000$. As for d, the number of components is 12 according to 12 dimensional feature vectors used in the following experiments. Considering these distributions and that in Bayesian parameter estimation for Gaussian distribution, Gaussian distribution is used as conjugate prior for mean and gamma distribution is used as that for inverse variance [12], we describe the distribution of $m_{k',d}^{c_i} - m_{k,d}$ as a Gaussian distribution with zero mean and that of $\sigma_{k,d}/\sigma_{k',d}^{c_i}$ as a gamma distribution.

Model parameters of these Gaussian and gamma distributions are estimated using the above-mentioned 80,000 samples for each component $d = 1, \cdots, 12$, and estimated parameters for odd number components are shown in Table 1. In the table, λ is the standard deviation parameter for Gaussian distribution and θ_1 and θ_2 are the parameters for gamma distribution described as

$$p(x) = \frac{1}{\theta_2^{\theta_1}\Gamma(\theta_1)} x^{\theta_1-1}\exp(-x/\theta_2), \tag{15}$$

where $\Gamma(\theta_1)$ is the gamma function. As for θ_1 and θ_2, the averaged values of those for all ds, $\theta_1 = 1.727$ and $\theta_2 = 0.526$ are used in the following experiments because of not very strong dependency on d. In fact, as is shown in Table 2 in Sect. 5, SV performance using artificial cohorts generated with these averaged θ_1 and θ_2 is almost the same as that with d-dependent ones.

Using these estimated parameters and given parameters $m_{k,d}$ and $\sigma_{k,d}$ for a claimed speaker, the parameters $m_{k,d}^{c_i}$ and $\sigma_{k,d}^{c_i}$ for its artificial cohorts $c_i, i = 1, \cdots, N$ are set as

$$m_{k,d}^{c_i} = m_{k,d} + \lambda_d u_{k,d}^{c_i}, \tag{16}$$

$$\sigma_{k,d}^{c_i} = \sigma_{k,d}/v_{k,d}^{c_i}, \tag{17}$$

where λ_d is the above λ parameter for the d-th component, $u_{k,d}^{c_i}$ is a random number from the standard normal distribution (Gaussian with zero mean and unit variance), and $v_{k,d}^{c_i}$ is a random number from the gamma distribution with $\theta_1 = 1.727$ and $\theta_2 = 0.526$. The distribution of $m_{k',d}^{c_i} - m_{k,d}$ and that of $\sigma_{k,d}/\sigma_{k',d}^{c_i}$ for artificial cohorts, which are derived by the same procedure as aforementioned one for real cohorts, are also shown in Fig. 1 and Fig. 2, respectively. As for difference of means $m_{k',d}^{c_i} - m_{k,d}$, the distribution for artificial cohorts is close to that for real cohorts. However, as for ratio of variances $\sigma_{k,d}/\sigma_{k',d}^{c_i}$, considerable difference exists between the two distributions.

With regard to the mixing coefficient a_k of the k-th Gaussian, we take a similar approach to that for $m_{k,d}$. First, the data $a_{k'}^{c_i} - a_k$ for all speakers are collected, where a_k and $a_{k'}^{c_i}$ are the mixing coefficient of k-th Gaussian for a claimed speaker and that of the corresponding k'-th Gaussian for its real cohorts $c_i, i = 1, \cdots, N$, respectively. Then we describe the distribution of $a_{k'}^{c_i} - a_k$ as a Gaussian distribution with zero mean. Using the estimated standard deviation parameter $\lambda_a = 0.040$ for Gaussian distribution and a given a_k for a claimed speaker, the parameters $a_k^{c_i}$ for its artificial cohorts $c_i, i = 1, \cdots, N$ are set as

$$a_k^{c_i} = \frac{a_k + \lambda_a u_k^{c_i}}{\sum_{k=1}^{K}(a_k + \lambda_a u_k^{c_i})}, \tag{18}$$

where $u_k^{c_i}$ is a random number from the standard normal distribution. However, effect of changing a_k on performance improvement was very small or almost nothing. For example, in the following experiment using 50 frames shown in Table 4, error rates with and without changing a_k was 5.52% and 5.54%, respectively.

5. Speaker Verification Experiments

For SV experiments, telephone speech data-set was used, which consists of isolated uttered Japanese 20 words produced two repetitions by 100 male speakers in two sessions spaced three to four months apart [13]. The speech data was low-pass filtered at 4.5 kHz and digitized at 10 kHz sampling rate. The digitized speech was pre-emphasized with

Table 2 Equal error rates (%) in SV experiments with different numbers of frames using artificial cohorts generated with two kinds of parameters θ_1 and θ_2: d-dependent parameters and the averaged ones.

	the number of frames			
	50	100	150	200
d-dependent parameters	5.58	3.57	3.11	2.86
averaged parameters	5.52	3.58	3.17	2.79

a first-order adaptive filter and subjected to 12th order LPC analysis with 25.6 msec Hamming window and 12.8 msec frame rate. In fact, the selective LPC analysis was applied to use the spectral information up to 4 kHz considering that the speech data is telephone speech. The twelve LPC cepstral coefficients obtained by this analysis were used as a feature vector for each time frame.

The data-set was divided into two sets: one set consists of first-uttered 20 words in two sessions and the other consists of second-uttered ones in two sessions. The former set was used for training and the latter set for test. For SV experiments, word utterances of each speaker are connected and used in an endless way. The number of tests per speaker is 100 for utterances of the same speaker and 100 for those of impostors, i.e., 10,000 in total for both cases. In each test, starting point of input is randomly selected and impostors are also randomly selected from 99 speakers excluding the relevant true speaker. Experiments with 20,000 tests are carried out five times, where different artificial cohorts are generated for each experiment. SV performance is measured by average of equal error rates (EERs) for five experiments.

In the following experiments, the covariance matrix $\mathbf{\Sigma}_k$ of each Gaussian distribution $g_k(\mathbf{y}_t; \mathbf{m}_k, \mathbf{\Sigma}_k)$ is assumed to be diagonal. The number of mixtures for GMM is 16, and the number of cohorts used are 50 for both real and artificial cohorts, both of which follow [6]. Given the likelihood for a claimed speaker $p_s(\mathbf{Y})$ and that for an other speaker among 99 speakers $p_i(\mathbf{Y})$, the speaker i is selected as a member of 50 cohorts for the claimed speaker if the difference of likelihood $|p_s(\mathbf{Y}) - p_i(\mathbf{Y})|$ is within the smallest 50 among 99. The universal background model, which is used as a conventional method in the following experiments, was here estimated by using training data from all 100 speakers.

With regard to the parameters θ_1 and θ_2 for gamma distribution in Eq. (15), SV experiments are carried out using artificial cohorts generated with two kinds of parameters θ_1 and θ_2: d-dependent parameters and the averaged ones of those for all ds. Experimental results are shown in Table 2. Both of two parameter settings produce almost same SV performance.

The proposed SV method using artificial cohorts is evaluated by comparing it with the previous method as well as several conventional methods: methods using the universal background model, real cohorts from the data-set, and a background model in [5], which we call pseudo background model. For reference, SV experiments using a method without background model were also carried out where only the log-likelihood for a claimed speaker $\log p_s(\mathbf{Y})$ is used instead of the log-likelihood ratio $S(\mathbf{Y})$ in (1). Regarding the method using pseudo background model [5], we have found out that a single Gaussian model instead of mixture models is best for the pseudo background model, which is shown in Table 3.

Table 3 Equal error rates (EERs) in SV experiments using several pseudo background models with different numbers of mixtures, where the number of frames used is 50.

	the number of mixtures			
	8	4	2	1
EER (%)	24.9	13.4	8.8	7.2

Table 4 Equal error rates (%) for several SV methods with different numbers of frames.

	the number of frames			
method	50	100	150	200
universal background	3.3	1.9	1.5	1.4
real cohort	2.5	1.5	1.4	1.4
artificial cohort (previous)	6.4	4.1	3.4	3.1
artificial cohort (proposed)	5.5	3.6	3.2	2.8
pseudo background	7.2	4.1	3.2	2.8
without background	7.6	5.4	4.8	4.5

Experimental results are shown in Table 4. SV performance by the method without background model is naturally worst among all methods. EERs for the proposed method using artificial cohorts are about 10% less than those for the previous method, and EERs for the proposed method are less than those for the method using pseudo background model when the number of frames used is small. However, SV performance by the proposed method using artificial cohorts is still much worse than that using real cohorts.

6. Conclusions

This paper discussed how to generate artificial cohorts in order to realize GMM-based SV method using only utterances of enrolled speakers, i.e., without using speaker databases including many other speakers' utterances. In this paper, we proposed a rational approach to set GMM parameters for artificial cohorts based on statistics of GMM parameters for real cohorts. EERs for the proposed method using artificial cohorts are about 10% less than those for the previous method, where GMM parameters for artificial cohorts were set in an ad hoc manner.

References

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digit. Signal Process., vol.10, no.1-3, pp.19–41, 2000.

[2] A.E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F.K. Soong, "The use of cohort normalized scores for speaker verification," Proc. ICSLP-92, pp.599–602, 1992.

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," Digit. Signal Process., vol.10, no.1-3, pp.42–54, 2000.

[4] O. Siohan, C.-H. Lee, A.C. Surendran, and Q. Li, "Background model design for flexible and portable speaker verification systems," Proc. ICASSP-99, pp.825–828, 1999.

[5] D. Tran and D. Sharma, "New background speaker models and experiments on the ANDOSL speech corpus," Lect. Notes Comput. Sci., vol.3214, pp.498–503, 2004.
[6] Y. Mukai, H. Noda, M. Niimi, and T. Osanai, "Text-independent speaker verification using artificially generated GMMs for cohorts," IEICE Trans. Inf. & Syst., vol.E91-D, no.10, pp.2536–2539, Oct. 2008.
[7] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Roy. Stat. Soc., vol.39, no.1, pp.1–38, 1977.
[8] G.J. McLachlan and K.E. Basford, Mixture models — Inference and applications to clustering, Marcel Dekker, 1988.
[9] 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.
[10] Y. Lind, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol.CON–28, no.1, pp.84–95, 1980.
[11] C.W. Therrien, Decision Estimation and Classification — An Introduction to Pattern Recognition and Related Topics, John Wiley & Sons, 1989.
[12] D.J.C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003.
[13] H. Noda, K. Harada, and E. Kawaguchi, "A context-dependent sequential decision for speaker verification," IEICE Trans. Inf. & Syst., vol.E82-D, no.10, pp.1433–1436, Oct. 1999.