情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

局所MFCC特徴量の最適線形結合による テキスト独立型話者照合

Text-Independent Speaker Verification using Optimized Linear Combination of Local MFCC Features

坂井俊亮* 亀山啓輔†
Shunsuke Sakai Keisuke Kameyama

Abstract: In recent years, studies of speaker verification have been conducted as a means for biometric person authentication. However, because of the overall verification performance, only few actual implementations exist. This paper focuses on the text-independent speaker verification system. We propose an effective method for speaker verification by adaptive weighting of local Mel Frequency Cepstrum Coefficient (MFCC) features. For a given set of registered persons, an optimal linear weightings of multiple speech frames are searched based on the likelihood ratio error, generalizing the scheme of the conventional use of Δ parameters [1]. It was observed that using the proposed adaptive parameters, superior verification performance was achieved compared with the cases using conventional features.

Keywords: text-independent speaker verification, biometric authentication, adaptive feature weighting, inter-frame feature.

1 Introduction

In recent years, as a technology to verify individuals, studies for authentication using human biometrics have been conducted actively [2]. In password authentication conventionally often used, there are problems that the users forget the phrase and impostor can be easily verified due to leakage or theft. Therefore, biometrics technology verifying the individuals using physical information such as fingerprint, vein pattern, iris, face, and speech is in the spotlight [3]. Speaker verification technology to verify the speaker by speech features can be useful as it does not require special verification hardware, is less stressful for the users, and can be used from remote places across the telephone network. However, actual use of speaker verification technology is still less common. The main reason is due to the fact that the verification performance is still low when compared with the use of other modalities.

This study proposes a framework of speaker modeling to use the inter-frame dynamics in addition to per-frame Mel Frequency Cepstrum Coefficient (MFCC) speech feature aiming to improve the verification precision. While ΔMFCC feature [1] is known to take a similar approach, it uses fixed weights of local MFCC features. In contrast, the proposed feature employs adaptive weights optimized for improving the verification performance. Also, some methods that present mel-cepstral analysis method and its adaptive algorithm [4], and method which optimizes the weights for each likelihood such that the overall expected loss can be minimized [5], have been reported. In this work, we examine the improvement of speaker discrimination ability by using an optimized linear combination

*筑波大学 システム情報工学研究科, 305-8573 茨城県つくば市天王台 1-1-1, tel. 029-853-6200 ext. 8480, e-mail sakai@adapt.cs.tsukuba.ac.jp,
Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
†筑波大学 システム情報工学研究科, 305-8573 茨城県つくば市天王台 1-1-1, tel. 029-853-6200 ext. 8480, e-mail Keisuke.Kameyama@cs.tsukuba.ac.jp,
Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

of relatively long-term features.

2 Text-independent speaker verification

The methods of speaker verification is divided into three main groups [6]. Text-dependent system specifies the verification text in beforehand, text-independent system does not limit the text, and text-prompted system prompts the text at each verification. In general, it is known that text-dependent and text-prompted systems indicate higher performance, because the system can use the speaker information depending on phonological line describing the text [7]. Recently, the study of text-independent speaker verification having an advantage of not limiting the speech text, has been the mainstream research topic [6]. This paper discusses text-independent verification as well.

2.1 Procedure

The general flow of text-independent speaker verification system is shown in Figs. 1 and 2 [6]. At first, in the modeling phase, the model of each speaker λ_C and the background model $\lambda_{\bar{C}}$ are obtained based on the speech signals. The speaker model λ_C uses a collection the authentic speech, whereas the background model (Universal Background Model [8]) uses speeches by various speakers (average feature) in the training. In the test phase, the log-likelihood ratio of input speech feature vectors to the claimed speaker model λ_C and the background model $\lambda_{\bar{C}}$ is calculated, and this value is compared to the predetermined threshold value θ. The speaker is accepted if the value is higher than the threshold value, and is rejected otherwise. The log-likelihood ratio is defined as

$$\Lambda_C(X) = \log p(X \mid \lambda_C) - \log p(X \mid \lambda_{\bar{C}}). \tag{1}$$

Here, if $\boldsymbol{x}(i)$ is a feature vector in frame i $(i = 1, 2, ..., T)$, $p(X \mid \lambda_C)$ indicates the likelihood that speech $X = \{\boldsymbol{x}(1), ..., \boldsymbol{x}(T)\}$ is from the claimed speaker giving the model λ_C. On the other hand, $p(X \mid \lambda_{\bar{C}})$ indicates the likelihood that speech X is not from the claimed speaker.

Additionally, the likelihood that the model of the claimed speaker gives the input speech collection X is

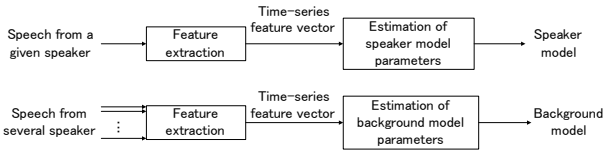

図 1: Modeling phase

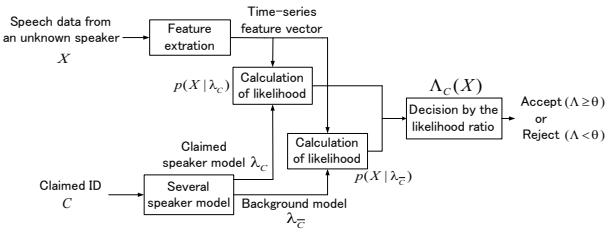

図 2: Test phase

defined as

$$\log p(X \mid \lambda_C) = \frac{1}{T} \sum_{t=1}^{T} \log p\left(\boldsymbol{x}(t) \mid \lambda_C\right). \tag{2}$$

As a speech feature, Mel Frequency Cepstrum Coefficient (MFCC) is used commonly [9]. Here, MFCC feature of D component cepstrum coefficients is described as $\boldsymbol{x} = [c_1, c_2, ..., c_D]'$.

2.2 Gaussian mixture speaker model

Text-independent speaker verification does not have a limitation about the speech contents by the speaker. Therefore, in the speaker modeling, likelihood function of speech feature is modeled as a density function, for example with Gaussian Mixture Model (GMM) [9]. Here, if $\boldsymbol{x}$ is a D-dimensional feature vector, likelihood function of a registered speaker s $(s = 1, ..., N)$ is defined as

$$p(\boldsymbol{x} \mid \lambda_s) = \sum_{j=1}^{M} w_{sj} b_{sj}(\boldsymbol{x}). \tag{3}$$

This is a linear combination of M Gaussian functions $b_{sj}(\boldsymbol{x})$, each computed as

$$b_{sj}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_{sj}|^{\frac{1}{2}}} \times \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_{sj})' (\Sigma_{sj})^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{sj}) \right\}, \tag{4}$$

which is determined by a mean vector $\boldsymbol{\mu}_{sj}$, a covariance matrix Σ_{sj}, and weights $w_{sj} (j = 1, ..., M)$. Here, the parameter set of speaker s is denoted as $\lambda_s = \{(w_{s1}, \boldsymbol{\mu}_{s1}, \Sigma_{s1}), ..., (w_{sM}, \boldsymbol{\mu}_{sM}, \Sigma_{sM})\}$. The speaker model parameters are estimated using the Expectation-Maximization (EM) algorithm [10].

3 Adaptive weighting of local MFCC features

It has been reported that, by adding inter-frame dynamic information to short-time per-frame speech feature as MFCC, there are cases that the verification performances are improved [1]. This feature known as the ΔMFCC regression coefficient of change along the time axis of MFCC, is sometimes used together with MFCC [6]. However, it only uses restrictive weights of local MFCC features.

We propose an adaptive method for weighting the local MFCC features that searches the optimal linear weightings of multiple speech frames based on the likelihood ratio error, generalizing the strategy taken by the Δ parameters employing restrictive weights.

3.1 Generalization of ΔMFCC feature

The inter-frame regression coefficient known as ΔMFCC is computed as

$$\Delta c_i(m) = \frac{\sum_{k=-l}^{l} k \cdot c_i(m+k)}{\sum_{k=-l}^{l} k^2}, \quad (i = 1, ..., D) \qquad (5)$$

where l is the frame range that regression coefficient is calculated, c is the cepstrum coefficient [1], and m is the frame index. In this work, we propose to generalize Eq. (5), searching for an arbitrary dynamic feature to improve verification precision among linear combinations of cepstrum coefficients in neighboring frames. Therefore, $2l+1$ parameters $\{a_s(-l), ..., a_s(l)\}$ in

$$c_{Fsi}(m) = \sum_{k=-l}^{l} a_s(k) c_i(m+k) \quad (i = 1, ..., D) \qquad (6)$$

are adjusted. Here, in the case of ΔMFCC, parameter $a_s(k)$ amounts to the special case of

$$a_\Delta(k) = \frac{k}{\sum_{k=-l}^{l} k^2}. \qquad (k = -l, ..., l) \qquad (7)$$

Figure 3 shows the schematic for calculating the feature c_{Fsi}.

3.2 Coefficient search based on likelihood ratio error

The verification precision can be evaluated by log-likelihood ratio. Therefore, in the proposed method, the coefficient parameter vector for each speaker

$$\boldsymbol{a}_s = [a_s(-l), a_s(-l+1), ..., a_s(l)]' \in \boldsymbol{R}^{2l+1} \qquad (8)$$

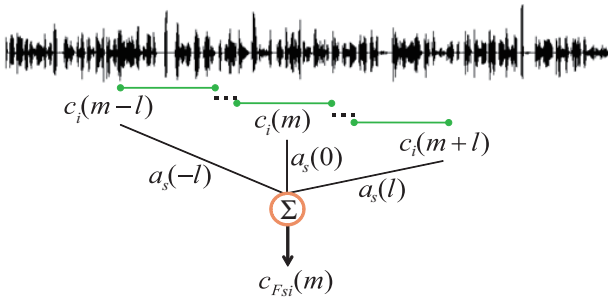

図 3: Extraction of weighted feature c_{Fsi}. Each horizontal bar denotes the period for a single frame.

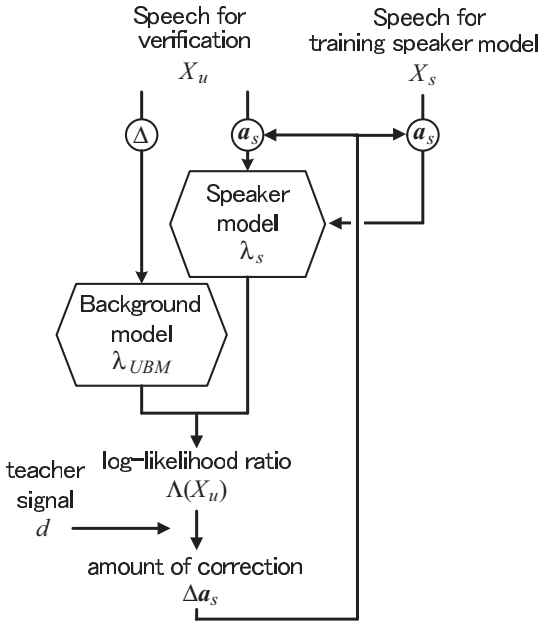

図 4: The training process. This modification is applied to the GMM parameter set λ_s and coefficient vector $\boldsymbol{a}_s$ of each registrant s $(s = 1, \ldots, N)$, iteratively.

is updated by steepest descent method to minimize the error in log-likelihood ratio for the teacher signal. Besides, each speaker has his/her own individual parameter vector. Figure 4 shows the whole procedure of parameter vector modification.

For a speech X_u attributed to speaker u, the T frames in X_u is converted to a feature vector array $\boldsymbol{g}_s(1), \ldots, \boldsymbol{g}_s(T) \in \boldsymbol{R}^{2D}$, using the parameters $\boldsymbol{a}_s$ and λ_s of a registrant s. The feature vector $\boldsymbol{g}_s(t)$ which concatenates MFCC features $\boldsymbol{x} = [c_1, c_2, ..., c_D]'$ and F-MFCC features $\boldsymbol{y}_s = [c_{Fs1}, c_{Fs2}, ..., c_{FsD}]'$, is denoted as

$$\boldsymbol{g}_s(t) = [\boldsymbol{x}', \boldsymbol{y}_s']' = M(t)\boldsymbol{a}_s + \boldsymbol{b}(t) \in \boldsymbol{R}^{2D}. \qquad (9)$$

$$M(t) = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \\ c_1(t-l) & \cdots & c_1(t+l) \\ \vdots & \ddots & \vdots \\ c_D(t-l) & \cdots & c_D(t+l) \end{bmatrix} \in \boldsymbol{R}^{2D\times(2l+1)} \quad \text{and} \tag{10}$$

$$\boldsymbol{b}(t) = [c_1(t), ..., c_D(t), 0, ..., 0]' \in \boldsymbol{R}^{2D}. \tag{11}$$

Then, the likelihoods for the background model and the corresponding speaker model are calculated by Eq. (3), and the log-likelihood ratio $\Lambda(X_u)$ is obtained as Eq. (1). Additionally, $\Lambda(X_u)$ is converted to a value in (0, 1) by the sigmoid function defined as

$$\sigma(\Lambda) = \frac{1}{1+\exp(-\beta\cdot\Lambda)} \tag{12}$$

considering the convergence performance in the next training phase. Here, β is the slope parameter. The initial $\boldsymbol{a}_s$ will be set identical to ΔMFCC coefficients $\{a_\Delta(-l), ..., a_\Delta(l)\}$.

Next, the coefficient parameter vector $\boldsymbol{a}_s$ is modified to minimize the error between the log-likelihood ratio $\sigma(\Lambda)$ and teacher signal. The teacher signal d is set as,

$$d = \begin{cases} 1 & (s=u) \ \text{(Authentic speaker)} \\ 0 & (s\neq u) \ \text{(Impostor)).} \end{cases} \tag{13}$$

Vector $\boldsymbol{a}_s$ is updated as

$$\boldsymbol{a}_s^{(\tau+1)} = \boldsymbol{a}_s^{(\tau)} + \Delta\boldsymbol{a}_s, \tag{14}$$

by the amount of correction defined as

$$\Delta\boldsymbol{a}_s = -\eta\frac{\partial E_{LLR}}{\partial \boldsymbol{a}_s} = -\eta\frac{\partial E_{LLR}}{\partial\sigma(\Lambda)}\frac{\partial\sigma(\Lambda)}{\partial\boldsymbol{a}_s}, \tag{15}$$

where, η is the learning coefficient. The log-likelihood ratio error E_{LLR} is defined as

$$E_{LLR} = \frac{1}{2}(d-\sigma(\Lambda))^2. \tag{16}$$

Each partial differentiation is computed as follows based on the definition above.

$$\frac{\partial E_{LLR}}{\partial\sigma(\Lambda)} = \frac{\partial}{\partial\sigma(\Lambda)}\frac{1}{2}(d-\sigma(\Lambda))^2 = -(d-\sigma(\Lambda)), \tag{17}$$

$$\frac{\partial\sigma(\Lambda)}{\partial\boldsymbol{a}_s} = \sum_{t=1}^{T}\left(\frac{\partial\sigma(\Lambda)}{\partial\boldsymbol{g}_s(t)}\frac{\partial\boldsymbol{g}_s(t)}{\partial\boldsymbol{a}_s}\right), \tag{18}$$

$$\begin{aligned}\frac{\partial\sigma(\Lambda)}{\partial\boldsymbol{g}_s(t)} &= \frac{\partial}{\partial\boldsymbol{g}_s(t)}\frac{1}{T}\sum_{i=1}^{T}\log p(\boldsymbol{g}_s(i)\mid\lambda_k) \\ &= \frac{1}{T}\frac{\partial}{\partial\boldsymbol{g}_s(t)}\log p(\boldsymbol{g}_s(t)\mid\lambda_k),\end{aligned} \tag{19}$$

$$\frac{\partial}{\partial\boldsymbol{g}_s(t)}\log p(\boldsymbol{g}_s(t)\mid\lambda_s) = -\frac{\sum_{j=1}^{M} c_{sj}(t)\{\boldsymbol{g}_s(t)-\boldsymbol{\mu}_{sj}\}'(\Sigma_{sj})^{-1}}{\sum_{j=1}^{M} c_{sj}(t)} \tag{20}$$

$$\text{where, } c_{sj}(t) = w_{sj}b_{sj}\left(\boldsymbol{g}_s(t)\right), \text{and} \tag{21}$$

$$\frac{\partial\boldsymbol{g}_s(t)}{\partial\boldsymbol{a}_s} = \frac{\partial}{\partial\boldsymbol{a}_s}\{M(t)\boldsymbol{a}_s + \boldsymbol{b}(t)\} = M(t). \tag{22}$$

Coefficient vector $\boldsymbol{a}_s$ for each registered speaker is updated using the amount of correction $\Delta\boldsymbol{a}_s$ in Eq. (15) by the following steps.

Training procedure

* Initialization

1) Train the background model (λ_{UBM}) and all the speaker model (λ_s) using feature vector (MFCC+ΔMFCC).

2) Calculate the log-likelihood ratio $\Lambda(X_u)$ using feature vector (MFCC+ΔMFCC) derived from speech X_u.

* Training of individual weight $\boldsymbol{a}_s$ and GMM

3) Calculate $\Delta\boldsymbol{a}_s$ to minimize the error E_{LLR} based on the teacher signal d, and update $\boldsymbol{a}_s$.

4) Train the speaker model (λ_s) using speech X_s again.

5) Verify using the new feature vector (MFCC+F-MFCC) for X_u by the updated model, and calculate the log-likelihood ratio.

6) Repeat 3-5 for the whole training set for a predetermined time.

4 Experiment

We evaluated the proposed method in four experiments. In Experiment 1, the optimal number of Gaussian component for our dataset is determined. In Experiment 2, the iteration number of learning $\boldsymbol{a}_s$ is determined. Experiment 3 compares the verification performance for the conventional method and the proposed method, and Experiment 4 is the same comparison using telephone speech.

4.1 Experimental condition

All speech data used in the experiments are sampled at 16kHz, in 16 bit. The speech data is extracted from the ASJ Continuous Speech Corpus by the Acoustical Society of Japan [11].

The number of clients are 10 (5 male, 5 female), and the speech length for speaker modeling is 60sec. The size of the training set is 100 trials (10 authentic, 90 impostor). The size of the test set is 10000 trials (500 authentic, 9500 impostor), with the speech length of 3sec. There are 50 speech clips from each speaker. The speech for the training and testing are from different texts. In the training and test phases, if one speaker is set to a client, all other speakers are set to impostors. The frame number l in Eq. (8) is set to $l = 2$, and the slope β in Eq. (12) is set to $\beta = 1.0$. All the experiments were conducted on a 3.6 GHz Intel Xeon computer with 3.0 GB of RAM, running Windows XP.

4.2 Experiment 1: Optimal number of Gaussian components

4.2.1 Procedure

In this experiment, we determined the optimal number of Gaussian component for our dataset. The background model and the speaker model with 16 to 512 Gaussian components were trained and evaluated using the same training and test data with clean (∞dB) speech. Equal Error Rate (EER) and the detection error tradeoff (DET) curve [12] are used as the index for evaluation.

4.2.2 Result

The comparison of each case is shown in Table 1 and Figure 5. 128 and 256 Gaussian components achieved the best among all in EER, and 128 appears slightly better than all others in DET curves. These results show that more Gaussian components are not always better.

From this result, we use 128 Gaussian components in the following experiments.

4.3 Experiment 2: Change in MSE with the iteration number of learning

4.3.1 Procedure

In order to determine the appropriate training iteration, the change of error E_{LLR} during training was invetigated. Mean-squared error (MSE) E_{LLR} for the training and test sets were recorded at each training round (100 trials). Normal speech in Experiment 3 and telephone speech in Experiment 4 are used, these include the same three cases of S/N as Experiment 3 and 4.

表 1: Exp.1 Comparison of EER by Gaussian component

Gaussian component	16	32	64	128	256	512
EER (%)	2.53	2.53	2.45	2.20	2.20	2.40

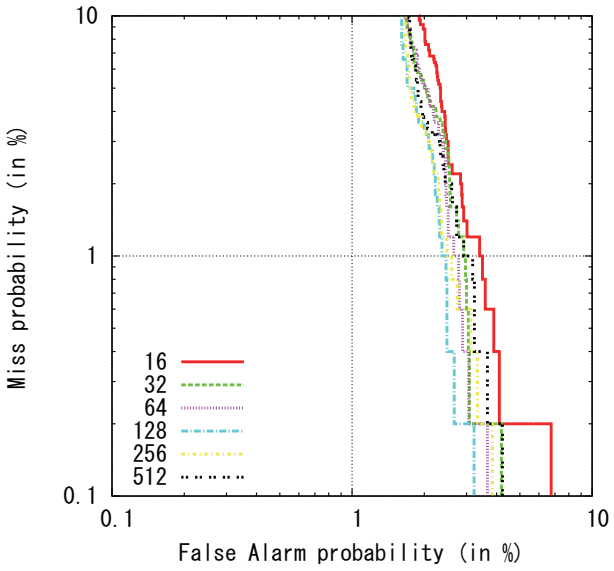

図 5: Exp.1 Comparison of DET curves by Gaussian component

4.3.2 Result

Figure 6 shows the case of normal speech in ∞dB test set and Figure 7 shows the case of telephone speech in ∞dB test set. These results show that the MSE of the test sets is minimum at the early stage in contrast to the monotonically decreasing training set error. The same tendency of these was seen in 30dB and 20dB test sets.

From this result, we set the iteration number of learning $\boldsymbol{a}_s$ varying GMM parameters at the same time to 400 (4 rounds of 100) in Experiment 3 and to 200 (2 rounds of 100) in Experiment 4. The value 400 was also used in Experiment 1.

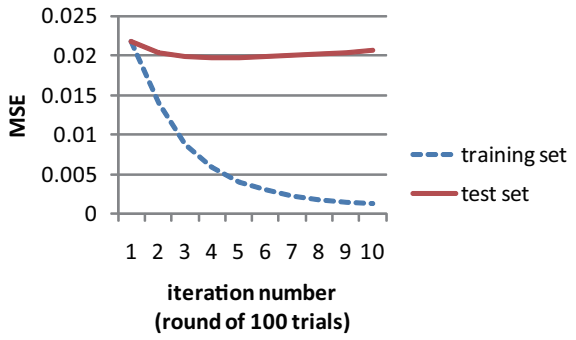

図 6: Exp.2 Change in MSE for training and test sets (normal, ∞dB)

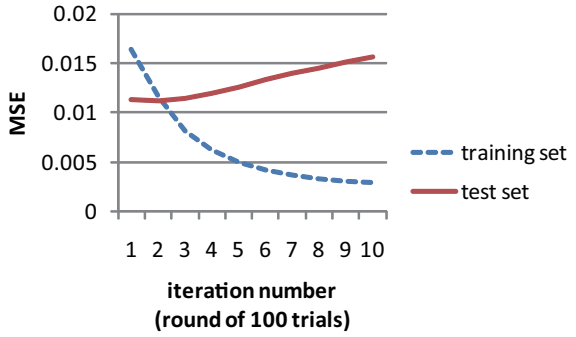

図 7: Exp.2 Change in MSE for training and test sets (phone, ∞dB)

4.4 Experiment 3: Comparison of the verification performance

4.4.1 Procedure

The conventional feature of MFCC, MFCC with ΔMFCC, and the proposed feature of MFCC with F-MFCC are compared in terms of the verification performance. The verification speech include three cases of S/N,
namely, ∞dB which is the ideal case, and levels common in usual verification use (30dB and 20dB). EER and DET curves are used as the index for evaluation as Experiment 1.

4.4.2 Result

The comparison using the EER(%) measure is shown in Table 2. Figs. 8-10 show the DET curves obtained for each S/N. In the case of training and testing with clean speech, the proposed method is superior to the conventional ones for verification speech with ∞dB. Additionally, in the case of training and testing with noisy speech, the proposed feature indicates higher performance than the conventional ones. Thus, the superiority of adaptively choosing the weighting of local MFCC features (MFCC+F-MFCC) has been verified.

表 2: Exp.3 Comparison of EER (%)

	∞dB	30dB	20dB
MFCC	2.80	2.60	2.20
MFCC+ΔMFCC	2.69	1.94	2.05
MFCC+F-MFCC	**2.20**	**1.89**	**1.34**

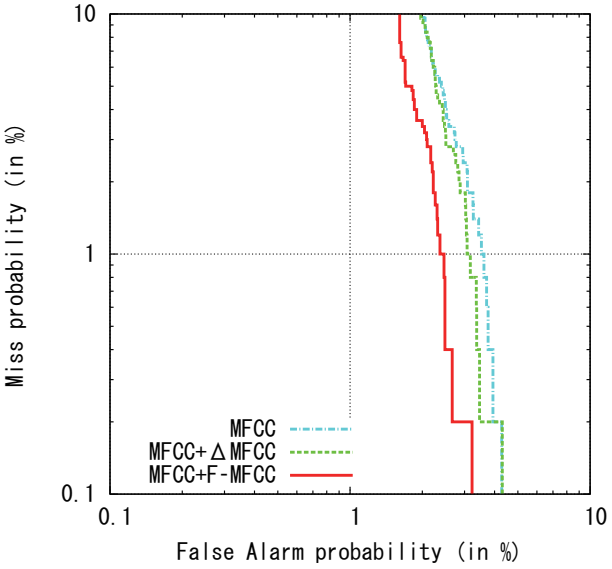

図 8: Exp.3 DET curves in ∞dB test set.

4.5 Experiment 4: Comparison of the verification performance (telephone speech)

4.5.1 Procedure

The same procedure as Experiment 3 is conducted with telephone speech. The frequency range of 300-3400 Hz was extracted from the same dataset, and was used as the simulated voice through a telephone line. The simulated telephone speech were used in both training and testing.

4.5.2 Result

The comparison using the EER(%) measure is shown in Table 3, and the DET curves are shown in Figs. 11-13. Here, the results also imply the superiority of using the proposed feature MFCC+F-MFCC in verification

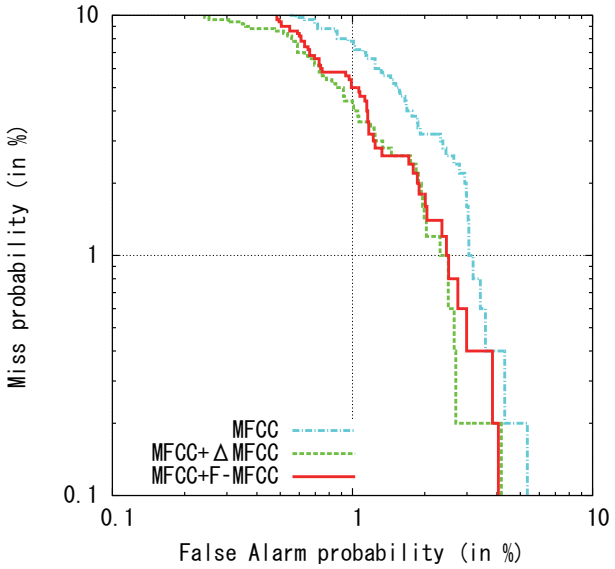

図 9: Exp.3 DET curves in 30dB test set.

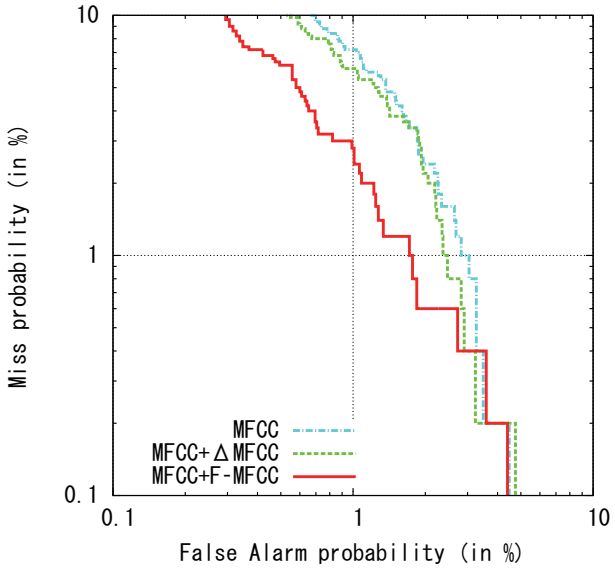

図 10: Exp.3 DET curves in 20dB test set.

表 3: Exp.4 Comparison of EER (%)

	∞dB	30dB	20dB
MFCC	1.80	3.20	2.80
MFCC+ΔMFCC	1.60	2.20	1.74
MFCC+F-MFCC	**1.40**	**1.46**	**1.40**

through the telephone line.

5 Conclusion

In this paper, we proposed a method for speaker verification using adaptive weighting of local MFCC features. The core idea was to determine the optimal frame coefficient parameter to minimize the verification error. In the experiments, it was shown that this mechanism gives lower verification error than the conventional methods under clean and certain level of noise environment, including authentication via a telephone line.

In future works, we consider methods that the present linear combination of local MFCC is extended to nonlinear transformation and indices other than the likelihood ratio error are used.

参考文献

[1] S. Furui, "Comparison of speaker recognition methods using static features and dynamic features," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 3, 1981.

[2] A. K. Jain, P. Flynn, and A. Ross, *Handbook of Biometrics*. Springer, 2008.

[3] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[4] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *ICASSP-92*, vol. 1, pp. 137–140, 1992.

[5] Y.-H. Chao, W.-H. Tsai, H.-M. Wang, and R.-C. Chang, "Improving the characterization of the alternative hypothesis via minimum verification error training with applications to speaker verification," *Pattern Recognition*, vol. 42, no. 7, pp. 1351–1360, 2009.

[6] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, 2004.

[7] T. Matsui and S. Kuroiwa, "Speaker recognition technology : A review and perspective," *Institute of Electronics, Information, and Communication Engineers*, vol. 87, no. 4, pp. 314–321, 2004.

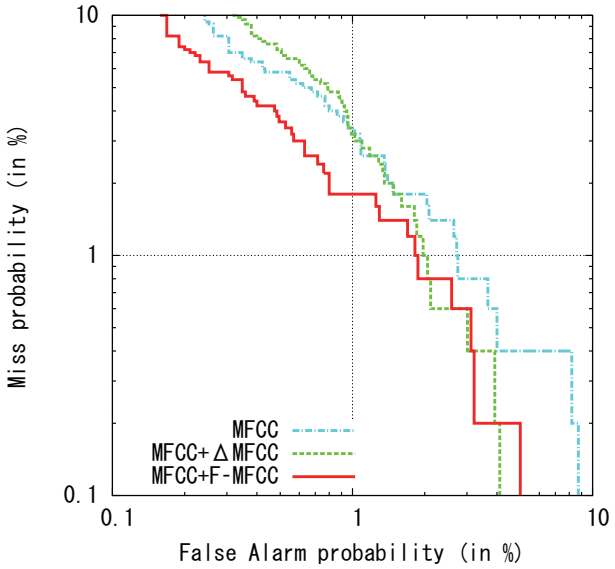

図 11: Exp.4 DET curves in ∞dB test set.

[8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, no. 10, pp. 19–41, 2000.

[9] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, no. 17, pp. 91–108, 1995.

[10] A.P.Dempster, N.M.Laird, and D.B.Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society (B)*, vol. 39, no. 1, pp. 1–38, 1977.

[11] S. Itahashi, M. Yamamoto, T. Takezawa, and T. Kobayashi, "Development of ASJ continuous speech corpus — Japanese newspaper article sentences (JNAS) —," *Proceedings of COCOSDA'97*, 1997.

[12] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," *Proceedings of European Conference on Speech Communication and Technology*, pp. 1895–1898, 1997.

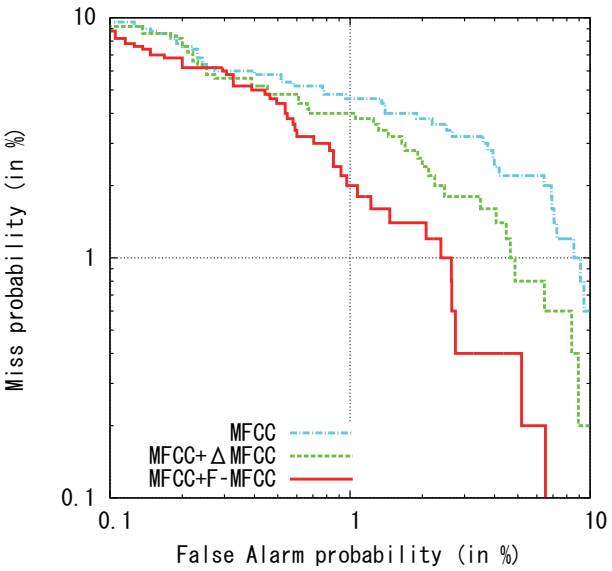

図 12: Exp.4 DET curves in 30dB test set.

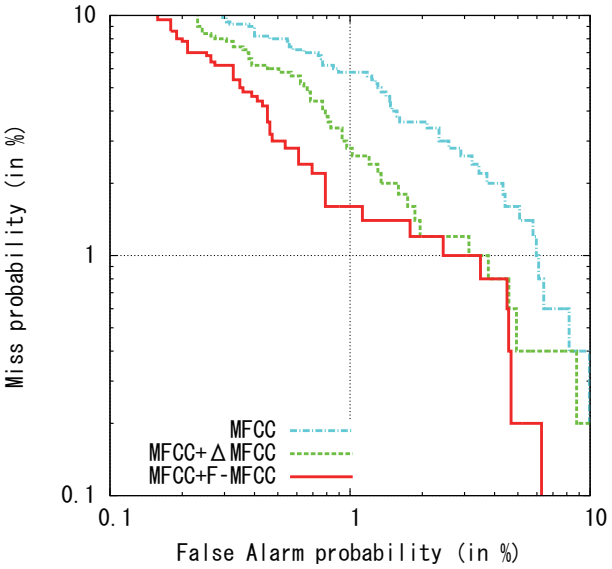

図 13: Exp.4 DET curves in 20dB test set.