PAPER

Text-Independent Speaker Identification utilizing Likelihood Normalization Technique

Konstantin MARKOV[†], ***Student Member and*** **Seiichi NAKAGAWA**[†], ***Member***

SUMMARY In this paper we describe a method, which allows the likelihood normalization technique, widely used for speaker verification, to be implemented in a text-independent speaker identification system. The essence of this method is to apply likelihood normalization at frame level instead of, as it is usually done, at utterance level. Every frame of the test utterance is inputed to all the reference models in parallel. In this procedure, for each frame, likelihoods from all the models are available, hence they can be normalized at every frame. A special kind of likelihood normalization, called *Weighting Models Rank*, is also experimented. We have implemented these techniques in speaker identification system based on VQ-distortion codebooks or Gaussian Mixture Models. Evaluation results showed that the frame level likelihood normalization technique gives higher speaker identification rates than the standard accumulated likelihood approach.

key words: *speaker identification, likelihood normalization, frame level processing*

1. Introduction

Speaker recognition has been research topic for many years and various types of speaker models have been studied. Hidden Markov Models (HMM) have become the most popular tool for this task. The best results have been obtained using Continuous HMM (CHMM) [2], [3]. Since temporal sequence modeling capability of the HMM is not essential for the text-independent task, one state CHMM, also called Gaussian Mixture Model (GMM), is widely used for speaker modeling [8], [9], [11], [12]. As our previous study [1] showed, GMM can perform even better than a CHMM with multi-states.

VQ-distortion codebook [3], [6] is another popular speaker model because of its non-parametric structure and its ability to model arbitrary data distributions. We used in our speaker identification system both the VQ-distortion codebook and GMM models and the standard accumulated distortion/likelihood testing serves as a baseline technique.

The likelihood normalization approach has been successfully applied for speaker verification [5], [9], [14], [15], but has never been used for speaker identification purposes. This is simply because when applied at utterance level, as in speaker verification, the likelihood normalization does not work [9]. In other words, when the reference speaker model scores (likelihoods) are calculated over the whole test utterance and then normalized, the likelihood normalization has no effect on the speaker identification rate. But Gish and Schmidt [11] have shown that when the speaker scores are computed over relatively short time intervals (segments of the utterances) likelihood normalization may be successful. In their system each speaker is represented by multiple GMMs trained on data from different sessions, and only the best model's score for each speaker over a given segment is taken into account. The scores are further normalized in order to obtain meaningful comparison between segments.

Our likelihood normalization approach makes use of a new speaker recognition system structure [17], [18], which is different from the study [11] in two main points. First, in our system each speaker is represented by only one GMM. Second, the speaker scores are computed at each frame instead of short time intervals. In other words, in our recognition system the test utterance is processed by all the reference speaker models in parallel in frame by frame manner. Having the likelihoods from all models, given particular test frame, allows these likelihoods to be normalized at the frame level. This frame level normalization is different from the standard normalization technique based on sentence level normalization and used for speaker verification [5], [14].Generally, the frame level likelihoods can be processed using not only normalization, but any appropriate technique such as transforming them into new scores. Transformed (normalized) likelihoods can further be accumulated over all test frames to form a final score for each speaker model. The identification is accomplished by identifying that speaker, whose model gives the best score.

2. Speaker models

In this section we give a brief description of the speaker models we used.

2.1 VQ-distortion codebook

Using LBG algorithm [16], from the each reference speaker training data a codebook model is trained such that the average distortion [6]:

Manuscript received October 1, 1996.
Manuscript revised January 16, 1997.

[†]The authors are with the Faculty of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi-shi, 441 Japan.

$$D = \frac{1}{T}\sum_{t=1}^{T} \min_{1\le j\le M} d(y_t, c_j) \tag{1}$$

is minimized over the whole training set $Y = y_1, \ldots, y_T$ and where $c_j, 1 \le j \le M$ are the centroids of the codebook and $d(,)$ is an Euclidean distance between two vectors.

The standard test approach consists of vector quantization of the test utterance $X = x_1, \ldots, x_T$ using all speakers codebooks and accumulation of the quantization errors (or distortions) with respect to each codebook across the whole test utterance. The average distortion with respect to the i^{th} codebook (speaker) is:

$$D^i = \frac{1}{T}\sum_{t=1}^{T} \min_{1\le j\le M} d(x_t, c_j^i) \tag{2}$$

and the final speaker identification decision is given by:

$$i^* = \arg \min_{1\le i\le N} D^i \tag{3}$$

where i^* is the identified speaker and N is the number of registered speakers.

2.2 Gaussian mixture model

A Gaussian mixture model is a weighted sum of M component densities and is given by the form [8]:

$$p(x|\lambda) = \sum_{i=1}^{M} c_i b_i(x) \tag{4}$$

where x is a d-dimensional random vector, $b_i(x)$, $i = 1, \ldots, M$, is the component density and c_i, $i = 1, \ldots, M$, is the mixture weight. Each component density is a d-variate Gaussian function of the form:

$$b_i(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)} \tag{5}$$

with mean vector μ_i and covariance matrix Σ_i. The mixture weights satisfy the constraint that:

$$\sum_{i=1}^{M} c_i = 1 \tag{6}$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation:

$$\lambda = \{c_i, \mu_i, \Sigma_i\} \quad i = 1, \ldots, M \tag{7}$$

In our speaker identification system, each speaker is represented by such GMM and is referred to by his/her model λ. GMM parameters are estimated using the standard Expectation Maximization (EM) algorithm.

For a sequence of T test vectors $X = x_1, \ldots, x_T$, the GMM log-likelihood can be written as:

$$L(X|\lambda) = \log P(X|\lambda) = \sum_{t=1}^{T} \log p(x_t|\lambda) \tag{8}$$

In the standard identification approach, the task is to find speaker whose model maximizes a posterior probability $P(\lambda_i|X), 1 \le i \le N$ which according to the Bayes rule is:

$$\begin{aligned} P(\lambda_i|X) &= \frac{P(X|\lambda_i)P(\lambda_i)}{P(X)} \\ &= \frac{P(X|\lambda_i)P(\lambda_i)}{\sum_{j=1}^{N} P(X|\lambda_j)P(\lambda_j)} \end{aligned} \tag{9}$$

Usually we don't have any prior knowledge about how likely unknown speaker is to be speaker i. That is why, the prior probabilities $P(\lambda_i)$ are assumed equal:

$$P(\lambda_i) = \frac{1}{N}, 1 \le i \le N \tag{10}$$

The term $P(X)$ is actually the probability of occurring the utterance X and in text-independent task is the same for all speakers. Therefore, $\max_{1\le i\le N} P(X|\lambda_i)$ will maximize the posterior probability and the identification decision can be simplified to:

$$i^* = \arg \max_{1\le i\le N} L(X|\lambda_i) \tag{11}$$

where i^* is the identified speaker.

3. The speaker identification system

Usually, speaker identification systems consist of collection of reference speaker models λ_i, front-end analysis and decision modules. Speech utterance is being transformed into a sequence of feature vectors X and after that the likelihoods $P(X|\lambda_i)$ (or distances D^i in case of VQ-codebook), corresponding to each of the speaker models, are calculated. The best one is determined in the decision module. This kind of speaker identification system allows only normalization of the final likelihoods $P(X|\lambda_i)$ or, in other words, only utterance (sentence) level likelihood normalization which is often used in speaker verification. In order to apply likelihood normalization at other level, for example frame level, the structure of identification system have to be modified.

Fig. 1 shows the structure of our speaker identification system [18]. In this system, input speech is analyzed and transformed into a feature vector sequence by Front-end Analysis block and then each test vector x_t is fed to all reference speaker models in parallel. The i^{th} speaker dependent GMM produces likelihood $p_i(x_t)$, $i = 1, 2, \ldots, N$ and all these likelihoods are passed in the so called *Likelihood processing* block, where they are transformed (normalized) and accumulated for $t = 1, 2, \ldots, T$ to form the new scores $Sc_i(X)$.

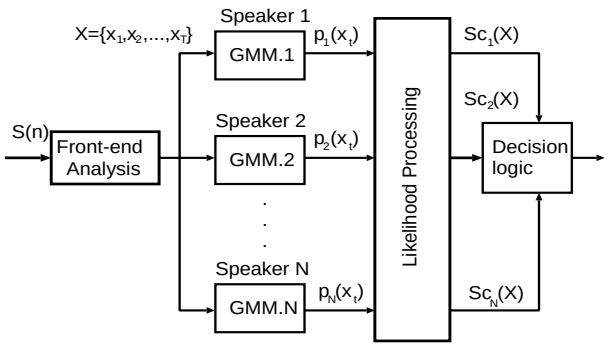


Fig. 1 Block diagram of the speaker identification system.

The speaker identification is accomplished by comparing these scores in the decision logic block and determining the best one . The unknown speaker is classified as the speaker, whose model has given the best score.

4. Frame level normalization

In this section we describe our approach to likelihood normalization at frame level and its implementation. A new method called Weighting models rank (WMR) is also presented. Both the likelihood normalization and WMR techniques change the value of frame likelihoods $p(x_t|\lambda_i)$, hence they can be called frame level likelihood transformation techniques.

4.1 Likelihood normalization

For speaker verification, likelihood normalization technique has been proved to improve significantly system performance [5], [9], [14]. The general approach is to apply a likelihood ratio test [13] to input utterance $X = x_1, x_2, \ldots, x_T$ using the claimed speaker model λ_c:

$$l(X) = \frac{P(\lambda_c|X)}{P(\lambda_{\overline{c}}|X)} \tag{12}$$

Applying Bayes' rule and assuming equal prior probabilities, the likelihood ratio becomes:

$$l(X) = \frac{P(X|\lambda_c)}{P(X|\lambda_{\overline{c}})} \tag{13}$$

where $\lambda_{\overline{c}}$ is a model representing all other possible speakers (impostors). The likelihood $P(X|\lambda_c)$ is directly computed from Eq.(8) assuming that the speaker model is of GMM type:

$$P(X|\lambda_c) = \prod_{t=1}^{T} p(x_t|\lambda_c) \tag{14}$$

The likelihood $P(X|\lambda_{\overline{c}})$ is usually approximated using a collection of *background* speaker models. With the set of B background speaker models, $\{\lambda_1, \ldots, \lambda_B\}$, the background speaker's likelihood is computed as:

$$P(X|\lambda_{\overline{c}}) = \frac{1}{B}\sum_{b=1}^{B} P(X|\lambda_b) \tag{15}$$

The likelihood normalization provided by the background speakers is important for the speaker verification task, because it helps to minimize the text dependent variations in the test utterance. The speaker identification task, based on utterance scores, does not need the normalization, because decisions are made using the likelihood from a single utterance requiring no inter-utterance likelihood comparisons [9].

But the situation for the speaker identification task becomes different when likelihood normalization is applied on the single vector likelihood $p(x_t|\lambda)$, or at the frame level. In this case, the likelihood normalization is done using:

$$p_{norm}(x_t|\lambda_i) = \frac{p(x_t|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B} p(x_t|\lambda_b)} \tag{16}$$

When B is big enough and approaches the number of reference speakers N, normalized likelihood $p_{norm}(x_t|\lambda_i)$ approximates a posterior probability $p(\lambda_i|x_t)$ because according to the Bayes rule:

$$\begin{aligned} p(\lambda_i|x_t) &= \frac{p(x_t|\lambda_i)p(\lambda_i)}{\sum_{j=1}^{N} p(x_t|\lambda_j)p(\lambda_j)} \\ &= \frac{p(x_t|\lambda_i)}{\sum_{j=1}^{N} p(x_t|\lambda_j)} \\ &\approx \frac{p(x_t|\lambda_i)}{\sum_{b=1}^{B} p(x_t|\lambda_b)} = \frac{1}{B} p_{norm}(x_t|\lambda_i) \end{aligned} \tag{17}$$

where a priori probabilities $p(\lambda_i)$ are assumed equal for all speakers. Therefore, this likelihood normalization is similar to normalization based on posteriori probability reported in [4]. However, it was applied for speaker verification and not at frame level.

In contrast to the speaker verification task, in speaker identification, there is no need of comparison of the normalized likelihoods with a threshold. Instead, they are accumulated over all vectors x_t, $t = 1, 2, \ldots, T$ for each speaker model i to produce the new scores:

$$Sc_i(X|\lambda_i) = \frac{1}{T}\sum_{t=1}^{T} \log p_{norm}(x_t|\lambda_i) \tag{18}$$

The speaker to be chosen, in this case, will simply depend on which speaker has the highest score $Sc_i(X|\lambda_i)$.

As in the speaker verification task, here also arises the problem of choosing the proper background speaker set. In the closed set speaker identification, the background speakers should be selected from the available set of N speakers. Given the speaker model i, the following background speaker sets seem to be reasonable:

- **All others** - the background speaker set consists

of all speakers, except the speaker i.

- **Top M speakers** - since the likelihoods from all speaker models for the current vector x_t are available, it is possible to determine the speaker models, which have the M maximum likelihoods and the background speaker set in this case consists of these M speakers (excluding speaker i). Obviously, the Top M speakers will change from frame to frame.
- **Cohort speakers** - the background speaker set consists of K acoustically most close speakers to the speaker i. The cohort speakers are determined on the training data in advance and this procedure is described in [14].

4.2 Weighting Models Rank (WMR)

This is a new technique which also transforms the frame likelihoods as does the likelihood normalization described above, but in rather different and deterministic way.

Since the likelihoods $p(x_t|\lambda_i)$ from all speaker models $i = 1, 2, \ldots, N$ for the current vector x_t are available, it is possible to sort them in order, corresponding to the value $p(x_t|\lambda_i)$. This is the same as to make N-best list of models for each vector x_t. At the top of this list is the model having highest likelihood and at the bottom, the model with the lowest likelihood. This procedure can be called also *ranking* of the speaker models. Table 1 shows how the speaker models are ordered in this list.

This table also shows that each rank (each row in the table) is assigned a weight $w_n, n = 1, 2, \ldots, N$. Now the scoring procedure is as follows:

- **Step 1**. For each test vector $x_t, t = 1, 2, \ldots, T$, construct the N-best list of the reference models $\lambda_i, i = 1, 2, \ldots, N$, as shown in the Table 1.
- **Step 2**. For each model $\lambda_i, i = 1, 2, \ldots, N$, find its rank n, i.e. its place in the N-best list, and assign the corresponding weight $w^i(t)$ to this model.
- **Step 3**. For each model λ_i, sum up all weights assigned to this model to produce its score:

$$Sc_i(X|\lambda_i) = \sum_{t=1}^{T} w^i(t) \tag{19}$$

where $w^i(t)$ is the weight of the model i at time t. The unknown speaker is identified as the speaker, who has the highest score $Sc(X|\lambda_i)$, i.e.:

$$i^* = \arg \max_{1 \le i \le N} Sc_i(X|\lambda_i) \tag{20}$$

Obviously, in this scoring approach, the most important issue is how to set the values of the weights w_n. We have found that weight values corresponding to exponential function shown in Fig. 2 give the best results

Table 1 N-best list of speaker models.

Rank r	Weight w_r	Model
1	w_1	Model λ_l (max.likelihood)
2	w_2	Model λ_j
...	...	...
m	w_m	Model λ_k
...	...	...
N	w_N	Model λ_p (min. likelihood)

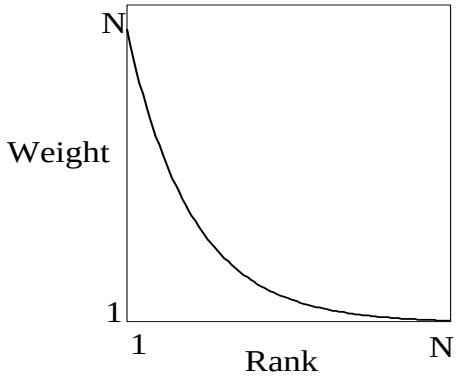


Fig. 2 Exponential weight function.

[18]. The exact values of the weights were calculated according to:

$$w_r = \exp\left(\frac{N - r + 1}{a}\right), r = 1, \ldots, N \tag{21}$$

where r is the current rank (see Table 1) and a is a scaling factor.

5. Databases and speech analysis

For the experiments we used two databases - NTT database and TIMIT corpus.

The NTT database consists of recordings of 35 speakers (22 males and 13 females) collected in 5 sessions over 10 months (1990.8, 1990.9, 1990.12, 1991.3 and 1991.6) in sound proof room [3]. For training the models, 5 same sentences for all speakers and 5 different sentences for each speaker, from one session (1990.8) were used. Five other sentences uttered at normal, fast and slow speeds and same for each of the speakers, from the other four sessions were used as test data. Average duration of the sentences is about 4 sec. The input speech was sampled at 12 kHz. 14 cepstrum coefficients were calculated by the 14th order LPC analysis at every 10 ms with a window of 21.33 ms. Then these coefficients were further transformed into 10 mel-cepstrum (cep) and 10 regressive (Δcep) coefficients. Each session's mel-cepstrum vectors were mean normalized by mean subtraction and silence parts were removed.

The well known TIMIT database, consisting of 6300 utterances (630 speakers$\times$10 utterances), was also used in evaluation experiments. 8 utterances (one SA, five SX and two SI) from each speaker were used for

Table 2 Identification rates (%) using GMM models (NTT database).

Model type	Feature	Normalization									Baseline		
		All others			Top 10			Cohort					
		Norm.	Slow	Fast	Norm.	Slow	Fast	Norm.	Slow	Fast	Norm.	Slow	Fast
4 mix. full	cep	92.8	89.0	90.9	92.7	89.2	90.7	92.4	89.4	91.2	92.3	88.6	90.4
	c+Δc	94.6	91.6	91.7	94.8	91.6	91.8	94.8	92.4	92.3	94.1	90.8	91.0
8 mix. full	cep	96.5	92.0	94.3	96.5	92.0	94.3	96.2	92.7	93.6	96.1	91.3	93.4
	c+Δc	97.0	93.4	94.6	97.0	93.6	94.5	97.0	93.8	94.3	97.0	93.0	94.0
32 mix. diag.	cep	95.5	92.7	92.6	95.5	92.7	92.6	95.2	92.6	93.2	95.0	92.4	91.7
	c+Δc	95.8	92.6	92.0	95.8	92.6	92.0	96.3	93.0	92.0	96.0	92.3	91.7
64 mix. diag.	cep	95.2	90.9	92.0	95.2	90.9	92.0	94.9	92.0	92.6	94.5	90.0	91.4
	c+Δc	95.7	91.6	92.3	95.7	91.5	92.3	95.9	91.7	91.9	95.4	91.0	91.4

training and the rest 2 (one SA and one SI) utterances for testing. The same speech analysis was performed as for the NTT database, except that cepstrum vectors were not mean normalized and silence was not removed.

6. Experiments

We evaluated our speaker recognition system using several types of GMMs with both full and diagonal covariance matrices and VQ-codebooks of different sizes. As a baseline system we used the standard approaches of Eq.(8,11) and Eq.(2,3).

6.1 NTT database results

The results presented in the following tables are averaged over all test sessions. Table 2 shows the identification rates using frame level likelihood normalization with the three types of background speaker set - All others, Top M with $M = 10$ and Cohort. Cohort size is set to $B = 5$. Three separate experiments were done for each type of the test utterances speeds (speaking rate) - normal, slow and fast. In the table, the columns marked with "Norm." ("N"), "Slow" ("S") and "Fast" ("F") show the identification rate in these three cases. Note that the speaker models were trained only with normal speed utterances. The column "Model type" shows the GMM structure. "4 mix. full" means a GMM with 4 mixture densities with full covariance matrices and "32 mix. diag." - GMM with 32 mixture densities with diagonal covariance matrices. The results in the "cep" rows present the identification rates when only 10-dimensional mel-cepstral feature vectors are used. Adding the cepstral derivative (Δcep) as a separate feature stream resulted in higher identification rates shown in the "c+Δc" rows. Table 2 shows that the frame level likelihood normalization gives better results than the baseline system. All types of the background speaker set give comparable identification rates. However, more important result is that the likelihood normalization technique is much better than the baseline at the "Slow" and "Fast" utterance speeds compared to the "Normal" speed. This fact shows that the frame level likelihood normalization approach is more robust against variations of the speaking rate.

Table 3 presents the results when Weighting models rank normalization technique is used. The exponential weights were ranging from $w_1 = 33.11, w_2 = 29.96, w_3 = 27.11\ldots$ to $w_{34} = 1.22, w_{35} = 1.10$. It is noted that identification rate of 97.3% is the best on this database (for comparison see [3]) and is achieved using WMR technique and GMM with 8 full covariance matrix mixtures.

When reference speakers are represented by the VQ-codebooks, likelihood normalization technique can be adapted interpreting the distortions as negative log-likelihoods, i.e:

$$d(x_t, C_i) = -\log p(x_t|\lambda_i) \tag{22}$$

where C_i is the i^{th} speaker codebook and λ_i is the corresponding GMM. From this equation follows that:

$$p(x_t|\lambda_i) = \exp(-d(x_t, C_i)) \tag{23}$$

Now using this relationship the normalization formulae for the VQ-distortions is:

$$d_{norm}(x_t, C_i) = \frac{\exp(-d(x_t, C_i))}{\frac{1}{B}\sum_{b=1}^{B}\exp(-d(x_t, C_b))} \tag{24}$$

where C_i is the current speaker's codebook and $C_b, b = 1, \ldots, B$ are the background speakers codebooks. The following accumulation of the normalized distortions is as in the case of likelihoods and the total score given the test utterance $X = x_1, \ldots, x_T$ is:

Table 3 Identification rates (%) using GMMs and weighting models rank normalization (NTT database).

Mod. type	Feature	WMR normalization			Baseline		
		N	S	F	N	S	F
4 m. full	cep	92.4	90.3	89.9	92.3	88.6	90.4
	c+Δc	95.2	91.0	91.9	94.1	90.8	91.0
8 m. full	cep	96.6	93.9	94.1	96.1	91.3	93.4
	c+Δc	97.3	94.3	94.8	97.0	93.0	94.0
32 m. diag.	cep	95.0	92.5	91.4	95.0	92.4	91.7
	c+Δc	95.3	92.6	90.5	96.0	92.3	91.7
64 m. diag.	cep	96.2	91.4	92.0	94.5	90.0	91.4
	c+Δc	95.8	91.9	92.4	95.4	91.0	91.4

Table 4 Identification rates (%) using VQ-codebook models (NTT database).

CB size	Feature	Likelihood normalization						WMR			Baseline		
		Top 10			Cohort								
		Norm.	Slow	Fast	Norm.	Slow	Fast	Norm.	Slow	Fast	Norm.	Slow	Fast
16	cep	86.4	82.8	81.6	85.8	82.7	82.0	86.6	84.7	81.3	85.3	82.6	80.4
	c+Δc	88.9	85.7	84.7	88.2	86.0	84.7	88.7	87.7	84.9	87.0	85.4	82.7
32	cep	93.0	91.2	89.4	92.7	91.2	89.2	92.4	91.7	87.3	91.0	89.7	87.0
	c+Δc	94.0	91.6	89.7	93.7	91.7	90.4	93.8	92.7	89.6	91.4	90.7	87.0
64	cep	92.2	91.8	90.3	92.2	92.0	90.3	93.3	92.0	89.7	91.4	92.0	89.4
	c+Δc	94.1	92.7	91.0	93.8	91.8	91.4	95.1	93.7	90.6	93.4	92.1	89.8
128	cep	94.4	93.0	91.3	94.2	93.0	91.3	94.0	93.1	91.0	94.1	93.2	91.3
	c+Δc	95.6	93.4	91.6	95.4	93.7	92.0	94.6	94.6	92.4	95.3	92.6	91.1

$$D^i_{norm} = \frac{1}{T}\sum_{t=1}^{T} \log d_{norm}(x_t, C_i) \tag{25}$$

$$= -D^i - \frac{1}{T}\sum_{t=1}^{T} \log(\frac{1}{B}\sum_{b=1}^{B} \exp(-d(x_t, C_b))) \tag{26}$$

where D^i is obtained form Eq.(2). In this case, the unknown speaker is identified by:

$$i^* = \arg \max_{1 \le i \le N} D^i_{norm} \tag{27}$$

Implementation of the WMR technique is easier because we need only to sort the distortions in reverse order. That is, the codebook with minimum $d(x_t, C_i), i = 1, \ldots, N$ is at the top of the table (see Table 1).

Table 4 shows the results of the experiments with VQ-codebook models and our normalization techniques. The column "CB size" shows the number of codewords in the codebook and the meaning of the other columns is the same as in Table 2. VQ-codebook models results show the same degree of superiority of the normalization techniques as in the GMMs case. Using the same train and test data identification rates of 90.9%, 93.0% and 93.9% (normal speed) for VQ-codebooks of sizes 32, 64 and 128 were reported in [3].

6.2 TIMIT database results

In Table 5, the results on TIMIT database are summarized. The column "Likelihood" means likelihood normalization using "All others" type of background speaker set (the other types are currently under experiments), and "WMR" means weighting models rank normalization with exponential weights. Identification rates for both the SA and SI test utterances are presented separately. Here also can be seen that our approaches give better results, except for the 4 mixture GMM.

Note that the best results for TIMIT database are achieved using WMR approach as for the NTT database, but this time using GMM with 16 full covariance matrix mixtures. The reason is that from the TIMIT data silence was not removed and thus, several of the GMM mixtures are necessary for modeling the silence. Since silences were removed from the NTT data, less mixtures were needed for the best performance.

Table 5 Identification rates (%) using GMMs (TIMIT database).

Mod. type	Feature	Normalization				Base line	
		Likelihood		WMR			
		SA	SI	SA	SI	SA	SI
4 m. full	cep	94.0	90.0	89.7	87.3	93.2	91.6
	c+Δc	94.8	91.1	89.8	87.0	95.1	92.9
8 m. full	cep	97.0	93.7	97.1	94.4	97.0	93.0
	c+Δc	97.3	94.1	95.7	93.0	96.8	93.8
16 m. full	cep	97.6	95.4	97.6	96.7	97.0	94.8
	c+Δc	96.8	93.8	98.1	95.1	96.7	94.4
16 m. diag.	cep	93.8	91.1	92.1	90.2	91.0	87.6
	c+Δc	94.1	90.8	89.4	86.3	92.4	87.9
32 m. diag.	cep	95.2	92.2	94.4	94.6	94.3	92.4
	c+Δc	94.9	92.1	94.1	91.4	94.3	92.4

7. Discussion

7.1 Statistical test of the improvements

In order to investigate the significance of the improvements achieved using frame level likelihood normalization and WMR techniques, we performed statistical significance test on the obtained results. For this, we used the sign test methodology described in [19].

When performing statistical significance test on the experimental results, the number of test samples by which the identification system was tested, is an important issue. For example, for few test samples, an acceptable significance level (or risk) will require quite big difference between identification rates of the standard and the new systems. In the case of NTT database, we had 5 test sentences per each of one of the 35 speakers which gives 175 tests per session, or 700 tests for all sessions. It is natural to take the best results of the both baseline and our system to calculate the significance level. Our baseline system achieved 97.0% identification rate for the 8 mixture full covariance matrix GMM. Having only 700 tests, a significance level of about 5% would require improvement of this result up to more than 98.6%. To achieve such identification rate is very

Table 6 Significance level (risk) (%) of the improvements for NTT database. (GMM - 8 mixture, full matrix).

Feature	WMR	Cohort
cep	0.1	7.3
cep+Δcep	4.4	11.9

difficult task on this database (it may not exist such a powerful method). However, since our test data consist of sentences uttered at normal, slow and fast speeds, we can perform significance test using averaged results of these three test conditions. This increases the number of the test samples to 2100. We performed significance test on the results obtained using 8 mixture full covariance matrix GMM, WMR technique and likelihood normalization with "Cohort" type background speaker set (this set performs best among the all sets). Table 6 shows the significance levels (risk) when both the "cep" and "cep+Δcep" feature vectors are used. These results confirm that WMR technique is best with significance level of 0.1% for "cep" features and 4.44% for "cep+Δcep" features. In other words, we can say with risk of only 4.4% that WMR is better than baseline for the NTT database when used with 8 mixture, full matrix GMM and cepstral and Δcepstral feature vectors, which gave the best result. The WMR identification rate for the TIMIT database (best performing configuration) also gives similar significance level.

7.2 Amount of computation

It is known that each improvement of any system increases its complexity. In our speaker identification system, since likelihood normalization is applied at frame level, for each test vector more computations are required than in the standard one. However, different background speaker sets require different number of additional operations. It is straightforward to estimate this number from the normalization formulae (Eq.(16)). For the "cohort" background speaker set two divisions and B summation ($2 * div. + B * sum.$) operations are required. Since the cohort size is a priori determined, in this case the number of additional operations does not depend on the number of reference speakers N and is roughly not more than 0.3% of the operations per frame for the baseline system using diagonal covariance matrices. If full covariance matrices are used additional amount of computation becomes negligible. "All others" background speaker set, however, requires $2*div.+(N-1)*sum$ and the amount of additional computation depends linearly on N. "Top M" background speaker set as well as WMR technique require frame likelihoods from all models to be sorted and in these cases the number of additional operations is a nonlinear function of N. When importance is given to the ratio performance/speed of the identification system, cohort frame likelihood normalization will be appropriate.

7.3 Frame level normalization

Here we would like to discuss about the normalization problem since it appears to be important to understand why likelihood normalization works only at the frame level for the speaker identification task.

First of all, the difference follows from the definitions of the frame and sentence level likelihood normalization, though the normalization formula is the same. For frame level likelihood normalization it is:

$$p^{norm}(x_t|\lambda_i) = \frac{p(x_t|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B} p(x_t|\lambda_b)} \tag{28}$$

where B is the number of background speakers. Sentence level likelihood normalization is defined as:

$$P^{norm}(X|\lambda_i) = \frac{P(X|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B} P(X|\lambda_b)} \tag{29}$$

From these two equations follows that the sentence log-score is:

$$\begin{aligned} Sc^{frame}(X|\lambda_i) &= \\ &= \sum_{t=1}^{T}(\log p(x_t|\lambda_i) - \log(\frac{1}{B}\sum_{b=1}^{B} p(x_t|\lambda_b))) \\ &= \sum_{t=1}^{T}\log p(x_t|\lambda_i) - \sum_{t=1}^{T}\log(\frac{1}{B}\sum_{b=1}^{B} p(x_t|\lambda_b)) \\ &= \log P(X|\lambda_i) - \log\prod_{t=1}^{T}(\frac{1}{B}\sum_{b=1}^{B} p(x_t|\lambda_b)) \end{aligned} \tag{30}$$

for the frame level normalization (actually, for the Top M background speaker set λ_b varies depending on the current frame and more correct would be to write $\lambda_b(t)$, but for the other sets background speakers are the same for all frames) and:

$$\begin{aligned} Sc^{sent}(X|\lambda_i) &= \\ &= \log P(X|\lambda_i) - \log(\frac{1}{B}\sum_{b=1}^{B} P(X|\lambda_b)) \\ &= \log P(X|\lambda_i) - \log(\sum_{b=1}^{B}(\frac{1}{B}\prod_{t=1}^{T} p(x_t|\lambda_b))) \end{aligned} \tag{31}$$

for the sentence level likelihood normalization. The final formulae became quite different. This shows that two normalization methods work in different way, but does not show why sentence level likelihood normalization has no effect on the identification rate. For this, each kind of the background speaker sets must be considered.

Top M - By definition this background speaker set is the same for all speaker models. That is why the normalization would be done by the value, which does not depend on the current speaker i and is the

same for all of them. If for any two speakers i and j the likelihoods are P_i and P_j, then the normalized likelihoods are:

$$P_i^{norm} = \frac{P_i}{\frac{1}{M}\sum_{b=1}^{M} P_b} = \frac{P_i}{C} \tag{32}$$

$$P_j^{norm} = \frac{P_j}{\frac{1}{M}\sum_{b=1}^{M} P_b} = \frac{P_j}{C} \tag{33}$$

Since the identification procedure is based on comparisons, the ratio between two likelihoods is important. If

$$\frac{P_i}{P_j} = k \tag{34}$$

then

$$\frac{P_i^{norm}}{P_j^{norm}} = \frac{\frac{P_i}{C}}{\frac{P_j}{C}} = \frac{P_i}{P_j} = k \tag{35}$$

which means that this kind of background speaker set does not change this ratio and, therefore, the identification rate, no matter at sentence or at frame level the normalization is applied.

That is why, a modified Top M background speaker set was used in the experiments. The modification consists of excluding the current speaker i from the background set in case his/her model's likelihood has been among the top M likelihoods. Thus, the modified Top M background speaker set changes the ratio (35) in a manner similar to all others background set, but only for the top M speakers at each frame.

All others - This background speaker set differs among the speakers. Following the same considerations, we have:

$$\begin{aligned} P_i^{norm} &= \frac{P_i}{\frac{1}{N-1}\sum_{b\neq i} P_b} \\ &= \frac{P_i}{\frac{1}{N-1}(\sum_{b=1}^{N} P_b - P_i)} \\ &= \frac{P_i}{C - \frac{P_i}{N-1}} \end{aligned} \tag{36}$$

The same holds for speaker j:

$$P_j^{norm} = \frac{P_j}{C - \frac{P_j}{N-1}} \tag{37}$$

Then if

$$\frac{P_i}{P_j} = k \tag{38}$$

the normalized likelihoods ratio becomes:

$$\begin{aligned} \frac{P_i^{norm}}{P_j^{norm}} &= \frac{P_i(C - \frac{P_j}{N-1})}{P_j(C - \frac{P_i}{N-1})} \\ &= k\frac{(N-1)C - P_j}{(N-1)C - kP_j} \\ &= k\frac{A - P_j}{A - kP_j} \end{aligned} \tag{39}$$

where $A = (N-1)C$. Now, if $k > 1$ then:

$$k\frac{A - P_j}{A - kP_j} > k \tag{40}$$

and if $k < 1$ then:

$$k\frac{A - P_j}{A - kP_j} < k \tag{41}$$

This means that this type of background speaker selection gains the ratio between likelihoods, but does not invert the inequality. Therefore, applied only once at sentence level it would not change the identification rate.

Cohort - With this type of background speaker set it is difficult to show analytically, that it does not work on sentence level normalization, because the cohort speakers are quite different for each speaker. However, it is obvious that it also changes the ratios between model likelihoods and, used in the frame level normalization, it gives improvement in the identification rate.

Weighing models rank - This technique, viewed as a special case of the likelihood normalization, can be applied only at the frame level. By its definition, it apparently changes proportions between model likelihoods, hence the identification rate. However, experiments showed that not every kind of the weight function gives the desired results.

8. Conclusion

We have experimented a new structure of the speaker identification system, which allows the likelihood normalization method to be utilized at frame level. A new technique, *Weighting Model Rank*, was also experimented. Both approaches showed better results in the speaker identification task compared to the standard accumulated likelihood/distortion methods on both the TIMIT and NTT databases. It was shown that any transformation (normalization) of the likelihoods at the frame level, which changes the ratio between them, influences the speaker identification rate.

We have confirmed that frame level normalization technique described here is also effective in speaker verification [17].

References

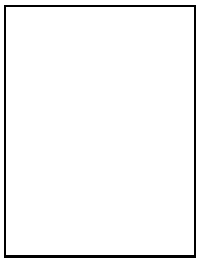
[1] K.Markov and S.Nakagawa, "Text-Independent speaker identification on TIMIT database", Proceedings, Acous.Soc.Jap. pp.83-84, March 1995.

[2] S. Furui, "Speaker-dependent feature extraction, recognition and processing techniques", Speech Communication, Vol.10, No. 5-6, pp. 505-520, 1991.

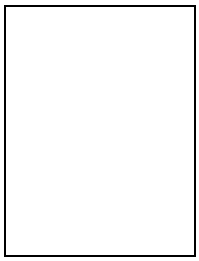
[3] T. Matsui and S. Furui, "Comparison of text independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs", Proc. ICASSP, Vol.II, pp.157-160, 1992.

[4] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition", Proc. ICASSP, pp.II-391-394, 1993.

[5] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model", Speech Communication, Vol. 17, No. 1-2, pp.109-116, 1995.
[6] F.K. Soong, A.E. Rosenberg, L.R. Rabiner & B.H. Juang, "A vector quantization approach to speaker recognition", AT&T Technical Journal, Vol. 66, pp.14-26, 1987.
[7] N.Z. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition", IEEE Trans. Signal Processing, Vol.39, pp.563-570, 1991.
[8] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans. on Speech and Audio Processing, Vol.3, No.1, pp.72-83, 1995.
[9] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, Vol. 17, No. 1-2, pp.91-108, 1995.
[10] B. Tseng, F. Soong and A. Rosenberg, "Continuous probabilistic acoustic map for speaker recognition", Proc. ICASSP, Vol.II, pp. 161-164, 1992.
[11] H. Gish and M. Schmidt, "Text-independent speaker identification", IEEE Signal Processing Magazine, October, pp.18-32, 1994.
[12] F.Bimbot, I. Magrin-Chagnolleau and L. Mathan, "Second-order statistical measures for text-independent speaker identification", Speech Communication, Vol. 17, No. 1-2, pp. 177-192, 1995,
[13] K. Fukunaga, "Introduction to statistical pattern recognition", Academic Press Inc., 1990.
[14] A. Rosenberg, J. DeLong, C. Lee, B. Juang and F. Soong, "The use of cohort normalized scores for speaker verification", Proc. ICSLP, pp.599-602, 1992.
[15] A. Higgins, L. Bahler and J. Porter, "Speaker verification using randomized phrase prompting", Digital Signal Processing, Vol. 1, pp. 89-106, 1991.
[16] Y. Linde, A. Buzo, and R.M. Gray. "An algorithm for vector quantizer design", IEEE Trans. Commun., Vol. COM-28, pp.84-95, 1980.
[17] K. Markov and S. Nakagawa, "Text-independent speaker recognition system using frame level likelihood processing", Technical report of IEICE, SP96-17, June 1996.
[18] K. Markov and S. Nakagawa, "Frame level likelihood normalization for text-independent speaker identification using Gaussian mixture models", Proc. ICSLP, pp.1764-1767, 1996.
[19] S. Nakagawa and H. Takagi, "Statistical methods for comparing pattern recognition algorithms and comments on evaluating speech recognition performance", The Journal of the Acoustical Society of Japan, Vol.50, No.10, pp.849-854, 1994.



Konstantin Markov was born in Bulgaria. He received his B.E. degree in Electrical Engineering from Department of Cybernetics, Leningrad Polytechnical Institute, Russia in 1984. He joined the Institute of Communication Industry, Sofia, Bulgaria in 1986 as a Research Associate. He received his M.E. degree in Electrical Engineering from Toyohashi University of Technology, Dep. of Information and Computer Sciences, Japan in 1996. He is currently a Ph.D. student at the Toyohashi University of Technology. His research interests include speaker identification/verification, language identification and pattern recognition.



Seiichi Nakagawa received his B.E. and M.E. degrees in Electrical Engineering from Kyoto Institute of Technology in 1971 and 1973, respectively, and his Dr. of Eng. degree from Kyoto University in 1977. He joined the Faculty of Kyoto University in 1976 as a Research Associate in the Department of Information Sciences. From 1980 to 1983, he was an Assistant Professor; from 1983 to 1990, he was an Associate Professor; and, since 1990, he has been a Professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi. From 1985 to 1986, he was a Visiting Scientist in the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, USA. He is the author of *Speech Recognition Based on Stochastic Model* (Inst. Elect. Inform. Comm. Engrs., Japan, 1988). Dr. Nakagawa was a co-recipient of the 1977 Paper Award from the IEICE and the 1988 J.C. Bose Memorial Award from the Institute of Electro. Telecomm. Engrs. His major interesting research areas are automatic speech recognition/speech processing, natural language processing, and artificial intelligence.