# FRAME LEVEL LIKELIHOOD NORMALIZATION FOR TEXT-INDEPENDENT SPEAKER IDENTIFICATION USING GAUSSIAN MIXTURE MODELS

*Konstantin MARKOV* *Seiichi NAKAGAWA*
*markov@slp.tutics.tut.ac.jp* *nakagawa@slp.tutics.tut.ac.jp*

Dept. of Information and Computer Sciences, Toyohashi Univ. of Tech.,
1-1 Hibarigaoka, Tempaku-chou, Toyohashi-shi, Aichi-ken, 441 JAPAN

## ABSTRACT

In this paper we propose a new speaker identification system, where the likelihood normalization technique, widely used for speaker verification, is introduced. In the new system, which is based on Gaussian Mixture Models, 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 proposed. Experiments were performed using two databases - TIMIT and NTT. Evaluation results clearly show that frame level likelihood normalization technique is superior to the standard accumulated likelihood approach.

## 1. INTRODUCTION

Speaker identification 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 [5, 6, 8, 9]. As our previous study [1] showed, GMM can perform even better than a CHMM with multi-states.

Although most of the existing speaker identification systems based on GMM address various problems, they have one thing in common. The reference speaker model scores (likelihoods) are calculated over the whole test utterance and then compared in order to find the best score. An exception is the system, studied by Gish and Schmidt [8], where the speaker scores are computed over relatively short time intervals (segments). In this 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 new speaker identification system structure, which is different from the study [8] 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 identification 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. Generally, the likelihoods can be processed using not only normalization, but any appropriate technique, which transforms them into a new scores. Transformed (normalized) likelihoods can further be accumulated over all test frames to form a final score for each speaker model. The unknown speaker is identified as the speaker, whose model gives the best score.

## 2. GAUSSIAN MIXTURE MODEL

A Gaussian mixture density is a weighted sum of $M$ component densities and is given by the form [5]:

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

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:

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

with mean vector $\mu_i$ and covariance matrix $\Sigma_i$. The mixture weights satisfy the constraint that:

$$\sum_{i=1}^{M} c_i = 1 \tag{3}$$

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

$$\lambda = \{c_i, \mu_i, \Sigma_i\} \qquad i = 1, \ldots, M \tag{4}$$

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

For a sequence of $T$ test vectors $X = x_1, x_2, \ldots, x_T$, the GMM log-likelihood can be written as:

$$P(X|\lambda) = \sum_{t=1}^{T} \log p(x_t|\lambda) \quad (5)$$

In the standard identification approach after applying the Bayes rule, the unknown speaker is identified from a set of $N$ speakers as:

$$i^* = \arg \max_{1 \le i \le N} P(X|\lambda_i) \quad (6)$$

# 3. SPEAKER IDENTIFICATION SYSTEM

Fig. 1 shows the structure of the new speaker identification system. The first step is, as usually, the transformation of the speech samples into a feature vector sequence $X = x_1, x_2, \ldots, x_T$. Then, each vector $x_t$ is fed to all reference speaker models in parallel, which is the main difference between this system and the standard one. The $i^{th}$ speaker dependent GMM produces the 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 $Sc_i(X)$. These scores are compared in the decision logic block and the best one is determined. The unknown speaker is classified as the speaker, whose model has given the best score.

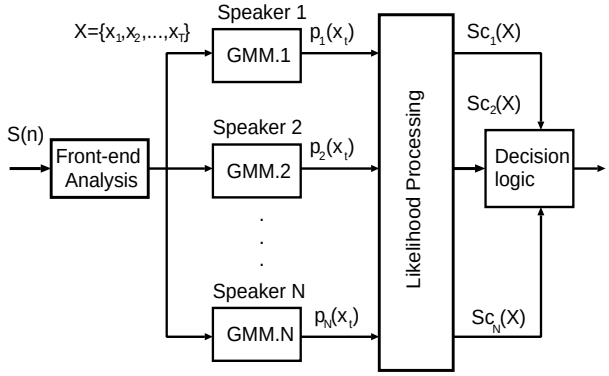


**Figure 1:** Block diagram of the new speaker identification system.

## 3.1. Likelihood Normalization

For the speaker verification, the likelihood normalization technique has been proved to improve significantly system performance [6, 11, 12]. The general approach is to apply a likelihood ratio test [10] to an input utterance $X = x_1, x_2, \ldots, x_T$ using the claimed speaker model $\lambda_c$:

$$l(X) = \frac{P(\lambda_c|X)}{P(\lambda_{\overline{c}}|X)} \quad (7)$$

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

$$\Lambda(X) = \log P(X|\lambda_c) - \log P(X|\lambda_{\overline{c}}) \quad (8)$$

where $\lambda_{\overline{c}}$ is a model representing all other possible speakers. The likelihood $P(X|\lambda_c)$ is directly computed from Eq.(5) assuming that the speaker model is of GMM type:

$$\log P(X|\lambda_c) = \frac{1}{T} \sum_{t=1}^{T} \log p(x_t|\lambda_c) \quad (9)$$

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 log-likelihood is computed as:

$$\log P(X|\lambda_{\overline{c}}) = \log \left\{ \frac{1}{B} \sum_{b=1}^{B} P(X|\lambda_b) \right\} \quad (10)$$

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 [6].

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:

$$p_{norm}(x_t|\lambda_i) = \frac{p(x_t|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B} p(x_t|\lambda_b)} \quad (11)$$

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:

$$Sc_i(X|\lambda_i) = \frac{1}{T} \sum_{t=1}^{T} \log p_{norm}(x_t|\lambda_i) \quad (12)$$

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 $i$.
- **All others from the same gender** - the background speaker set consists of all speakers having the same gender as speaker $i$, 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 $i$.

- **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 [11].

## 3.2. Weighting Models Rank

This is the new scoring approach and can be viewed as a special kind of likelihood normalization. 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, the model has the 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:

**Table 1:** N-best list of speaker models

| Rank | Weight | Model |
|---|---|---|
| 1 | $w_1$ | Model $\lambda_l$ (max.likelihood) |
| 2 | $w_2$ | Model $\lambda_j$ |
| ... | ... | ... |
| m | $w_m$ | Model $\lambda_k$ |
| ... | ... | ... |
| N | $w_N$ | Model $\lambda_p$ (min. likelihood) |

**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 $\lambda_i$, sum up all weights assigned to this model to produce its score:

$$Sc_i(X|\lambda_i) = \sum_{t=1}^{T} w^i(t) \tag{13}$$

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.:

$$i^* = \arg \max_{1 \le i \le N} Sc_i(X|\lambda_i) \tag{14}$$

Obviously, in this scoring approach, the most important issue is how to set the values of the weights $w_n$. Rather than to use any particular values for the weights, it seems to be reasonable to use values obtained according to a certain function. We used three types of functions as shown in Fig. 2.

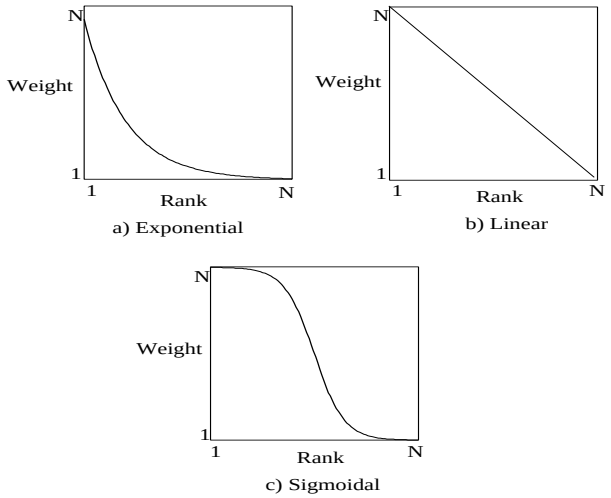


**Figure 2:** The three types of weight functions.

# 4. DATABASES AND SPEECH ANALYSIS

The NTT database consists of recordings of 35 speakers (22 males and 13 females) collected in 5 sessions over 10 months in sound proof room [3]. For training the models, 10 sentences from one session are used. Five other sentences 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 ($\Delta$cep) coefficients. Each session's mel-cepstrum vectors were mean normalized and silence parts were removed.

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

# 5. EXPERIMENTS

We evaluated our speaker identification system using several types of GMMs with both full and diagonal covariance matrices. As a baseline system we used the standard accumulated likelihood approach (Eq.(5)).

## 5.1. NTT 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 four types of

background speaker set. Cohort size is set to 5. Analyzing this table, it is clear that likelihood normalization gives better results and that Cohort background speaker set performs best. Table 3 presents the results when Weighting

**Table 2:** Identification rates (%) for GMMs using likelihood normalization.

| Model type | Feature | Backgr.speaker set | | | | Base line |
|---|---|---|---|---|---|---|
| | | All | Gen. | Top10 | Coh. | |
| 4 mix. full | cep | 95.02 | 95.02 | 95.02 | 95.30 | 95.02 |
| | c+$\Delta$c | 96.17 | 96.17 | 96.17 | 96.17 | 96.02 |
| 8 mix. full | cep | 96.30 | 96.15 | 96.30 | 96.15 | 95.87 |
| | c+$\Delta$c | 96.85 | 96.72 | 96.72 | 96.88 | 96.85 |
| 32 mix. diag. | cep | 96.15 | 96.00 | 96.15 | 96.15 | 96.00 |
| | c+$\Delta$c | 96.45 | 96.60 | 96.45 | 96.58 | 95.85 |
| 64 mix. diag. | cep | 96.57 | 96.57 | 96.73 | 96.85 | 96.15 |
| | c+$\Delta$c | 96.15 | 96.15 | 96.27 | 96.85 | 96.12 |

**Table 3:** Identification rates (%) for GMMs using weighting models rank normalization.

| Model type | Feature | Weight function | | | Base line |
|---|---|---|---|---|---|
| | | Sig. | Lin. | Exp. | |
| 4 mix. full | cep | 93.00 | 94.12 | 94.15 | 95.02 |
| | c+$\Delta$c | 92.60 | 94.72 | 95.55 | 96.02 |
| 8 mix. full | cep | 94.12 | 95.57 | 96.57 | 95.87 |
| | c+$\Delta$c | 95.00 | 96.70 | 97.85 | 96.85 |
| 32 mix. diag. | cep | 92.57 | 94.30 | 96.42 | 96.00 |
| | c+$\Delta$c | 93.00 | 94.55 | 96.42 | 95.85 |
| 64 mix. diag. | cep | 94.30 | 96.42 | 96.42 | 96.15 |
| | c+$\Delta$c | 95.27 | 95.97 | 97.15 | 96.12 |

models rank normalization technique is used with three types of weights. The importance of choosing the right weights is obvious. Only the exponential weights could outperform the baseline. It is noted that identification rate of 97.85% is the best on this database.

## 5.2. TIMIT results

In Table 4, 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, though the 4 mixture GMM did not perform well.

# 6. CONCLUSIONS

We proposed a new structure of the speaker identification system, which allows the likelihood normalization method to be utilized. A new technique, Weighting model rank, is also experimented. Both approaches showed better performance

**Table 4:** Identification rates (%) for GMMs using TIMIT database

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

compared to the standard accumulated likelihood on TIMIT and NTT databases.

# 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. 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.
5. 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.
6. D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, Vol. 17, No. 1-2, pp.91-108, 1995.
7. B. Tseng, F. Soong and A. Rosenberg, "Continuous probabilistic acoustic map for speaker recognition", Proc. ICASSP, Vol.II, pp. 161-164, 1992.
8. H. Gish and M. Schmidt, "Text-independent speaker identification", IEEE Signal Processing Magazine, October, pp.18-32, 1994.
9. 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,
10. K. Fukunaga, "Introduction to statistical pattern recognition", Academic Press Inc., 1990.
11. 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.
12. 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.