SPEAKER VERIFICATION USING FRAME AND UTTERANCE LEVEL LIKELIHOOD NORMALIZATION

Seiichi Nakagawa *Konstantin P. Markov*
Department of Information and Computer Sciences
Toyohashi University of Technology, Toyohashi, 441, Japan
{nakagawa,markov}@slp.tutics.tut.ac.jp

ABSTRACT

In this paper, we propose a new method, where the likelihood normalization technique is applied at both the frame and utterance levels. In this method based on Gaussian Mixture Models (GMM), every frame of the test utterance is inputed to the claimed and all background speaker models in parallel. In this procedure, for each frame, likelihoods from all the background models are available, hence they can be used for normalization of the claimed speaker likelihood at every frame. A special kind of likelihood normalization, called *Weighting Models Rank*, is also proposed. We have evaluated our method using two databases - TIMIT and NTT. Results show that the combination of frame and utterance level likelihood normalization in some cases reduces the equal error rate (EER) more than twice.

1. INTRODUCTION

Although most of the existing speaker verification systems based on GMM address various problems, they have one thing in common. The claimed speaker model score (likelihood) is calculated over the whole test utterance, normalized and then is compared with a threshold [1, 2, 3]. In other words, the likelihood normalization is done at the utterance level.

We have developed and implemented the frame level likelihood normalization method, firstly, for the speaker identification task and have shown its superiority over the standard accumulated likelihood approach [4]. For the speaker verification task, we first apply likelihood normalization at frame level and then at utterance level. For the frame level likelihood normalization we compute the frame likelihoods of the claimed speaker models as well as the background speaker models (a set of all registered speaker models). In other words, in our verification system the test utterance is processed by claimed and the background speaker models in parallel in frame by frame manner. Having the likelihoods from the background models, given particular test frame, allows the claimed speaker model's likelihood to be normalized at the frame level. Generally, at the frame level the claimed speaker model's likelihood can be processed using not only normalization, but any appropriate technique, which transforms it into a new scores. Transformed (normalized) likelihood is further accumulated over all test frames to form a final score for the claimed speaker model. Then, this score can be normalized again, at utterance level, and compared with a threshold as in the standard speaker verification systems.

Section 2. of this paper gives brief description of the GMM as a speaker model. In Section 3. we discuss the utterance and frame level likelihood normalization. A new normalization technique called *Weighting Models Rank (WMR)* is also presented. Section 4. introduces our speaker verification system. In Section 5. we describe the databases and the experimental results are summarized in Section 6..

2. GAUSSIAN MIXTURE MODEL

A Gaussian mixture density is a weighted sum of M component densities and is given by the form [3]:

$$p(x|\lambda) = \sum_{i=1}^{M} c_i N(x, \mu_i, \Sigma_i) \tag{1}$$

where x is a d-dimensional random vector, $b_i(x)$, $i = 1, \ldots, M$, is a Gaussian component density with mean μ_i and covariance matrix Σ_i, and c_i, $i = 1, \ldots, M$, is the mixture weight. 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: $\lambda = \{c_i, \mu_i, \Sigma_i\}$, $i = 1, \ldots, M$. In our speaker verification system, each registered speaker is represented by such a 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 $X = x_1, x_2, \ldots, x_T$, the GMM log-likelihood can be written as [6]:

$$\log P(X|\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(x_t|\lambda) \tag{2}$$

Then, the standard approach is to normalize $\log P(X|\lambda)$ and compare it with a threshold. If it exceeds the threshold, the claimed speaker is accepted, if not - it is rejected.

3. LIKELIHOOD NORMALIZATION

The essence of our method is to apply likelihood normalization at both the frame and utterance levels. This section describes it in more details.

3.1. At utterance level

For the speaker verification, the likelihood normalization technique has been proved to improve significantly system performance [2, 3, 7]. The general approach is to apply a likelihood ratio test [8] to an input utterance $X = x_1, x_2, \ldots, x_T$ using the claimed speaker model λ_c:

$$l(X) = \frac{P(\lambda_c|X)}{P(\lambda_{\overline{c}}|X)} \tag{3}$$

where $\lambda_{\overline{c}}$ is a model representing all other possible speakers. Applying Bayes' rule and assuming equal prior probabilities, the likelihood ratio in the log domain becomes:

$$\Lambda(X) = \log P(X|\lambda_c) - \log P(X|\lambda_{\overline{c}}) \tag{4}$$

The likelihood $\log P(X|\lambda_c)$ is directly computed from Eq.(2). The likelihood $\log P(X|\lambda_{\overline{c}})$ 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 [3]:

$$\log P(X|\lambda_{\overline{c}}) = \log \left\{ \frac{1}{B} \sum_{b=1}^{B} P(X|\lambda_b) \right\} \tag{5}$$

When the background speaker set consists of all registered speakers N, Eq.(3) becomes posteriori probability $P(\lambda_c|X)$ scaled by factor N:

$$\begin{aligned} l(X) &= \frac{P(X|\lambda_c)}{\frac{1}{N}\sum_{b=1}^{N} P(X|\lambda_b)} \\ &= \frac{P(X|\lambda_c)P(\lambda_c)}{\frac{1}{N}\sum_{b=1}^{N} P(X|\lambda_b)P(\lambda_b)} = NP(\lambda_c|X) \end{aligned} \tag{6}$$

3.2. At frame level

At the frame level, the likelihood normalization is applied on the single vector likelihood $p(x_t|\lambda)$. In this case, the likelihood normalization is done using:

$$p_{norm}(x_t|\lambda_i) = \frac{p(x_t|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B} p(x_t|\lambda_b)} \tag{7}$$

In contrast to the utterance level normalization, the normalized frame likelihoods are not compared with a threshold. Instead, they are accumulated over all vectors x_t, $t = 1, 2, \ldots, T$ to produce the new score:

$$Sc_i(X|\lambda_i) = \frac{1}{T} \sum_{t=1}^{T} \log p_{norm}(x_t|\lambda_i) \tag{8}$$

As in the utterance level normalization, here also arises the problem of choosing the proper background speaker set. Given the claimed speaker model i, we used the following background speaker sets [4]:

All others - the background speaker set consists of all registered speakers, except the speaker i.

Top M speakers - The likelihoods from all registered speaker models for the current vector x_t are computed, and those speaker models, which have the M maximum likelihoods are selected as the background speaker set (except the 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 [7].

3.3. Weighting Models Rank (WMR)

This is the new normalization approach where frame likelihoods are computed by all registered speaker models including the claimed speaker model. Then they are sorted 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. This procedure can be called also *ranking* of the speaker models. Table 1 shows how the speaker models are ordered in this list [4].

Table 1. N-best list of speaker models

Rank	Weight	Model
1	w_1	Model λ_l (max.likelihood)
...	...	...
m	w_m	Model λ_k
...	...	...
N	w_N	Model λ_p (min. likelihood)

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:

$$Sc_i(X|\lambda_i) = \sum_{t=1}^{T} w^i(t) \tag{9}$$

where $w^i(t)$ is the weight of the model i at time t.

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 typical function. We have found that for speaker identification, the exponential function shown in Fig.1, gives the best results [4] and here we used only this function.

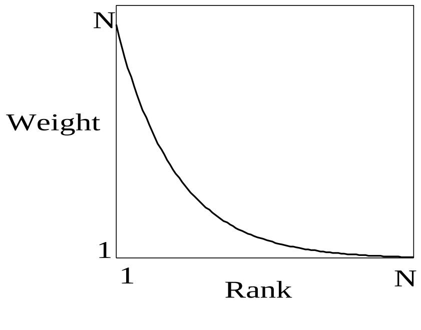

Figure 1. Weights as an exponential function of the model rank.

4. SPEAKER VERIFICATION SYSTEM

In order to have frame likelihoods from all background speakers available at each frame, a modification of the standard speaker verification system is necessary. Fig.2 shows our modified speaker verification system.

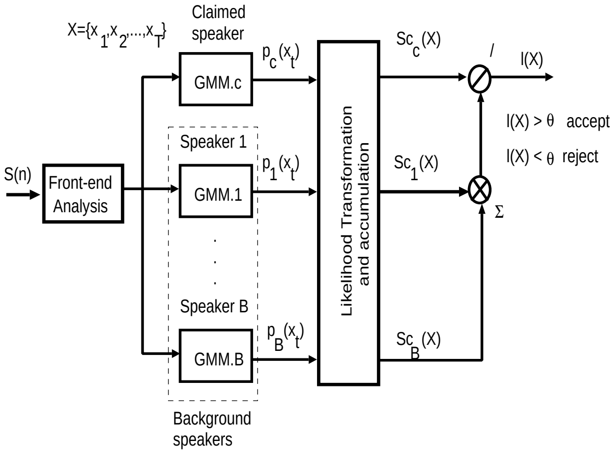

Figure 2. Block diagram of the modified speaker verification system.

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 the claimed speaker model GMM.c as well as to all background speaker models in parallel. The GMM.c gives likelihood $p_c(x_t)$ and background speaker models give $p_b(x_t), b = 1, \ldots, B$. All these likelihoods are passed to the so called *Likelihood transformation and accumulation* block, where they are normalized or transformed by WMR and accumulated for $t = 1, 2, \ldots, T$ to form the utterance level scores. The utterance scores $Sc_b(X), b = 1, \ldots, B$ are further used for utterance level likelihood normalization of the $Sc_c(X)$.

5. DATABASES AND SPEECH ANALYSIS

NTT database for speaker recognition consists of recordings of 35 speakers (22 males and 13 females) collected in 5 sessions over 10 months in sound proof room [9]. For training the models, 5 equal and 5 different sentences uttered at normal speed for each speaker from one session were used. Five other sentences uttered at normal, fast and slow speeds, 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 8 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 and silence parts were removed.

For the experiments on TIMIT corpus, 168 speakers (112 males and 56 females) from the "test" portion of the database were used. From the available 10 sentences per speaker 8 sentences (SAx1, SXx5 and SIx2) were used for training and 2 sentences (SA and SI) for testing. Speech data were analyzed using the same front-end parameters as for the NTT database, except that mel-cepstrum vectors were not mean normalized and silence parts were not removed. It has been found that cepstral mean normalization and silence removal of TIMIT data degrade the system performance [5].

6. EXPERIMENTS

In the experiments with NTT database each one of the 35 speakers was acting as customer (true test) while the others were used as impostors and verification was performed by rotating through all speakers and then averaging the results over all test sessions. This gives 175 true tests (35 × 5) and 5950 impostor tests (35 × 5 × 34) per session.

Results for the normal, slow and fast speed test utterances are reported in Tables 2, 3 and 4 respectively. The equal-error rate (EER) is computed a posteriori using a global threshold [2, 3]. In these tables "Background speakers" shows the type of the background speaker set used for frame likelihood normalization only. "All" means **All others** and "Coh." means **Cohort** types. Cohort consists of 5 speakers. For utterance level likelihood normalization we have tried the same types of background speaker sets and have obtained best EER with "Top 10" type. That is why only these results are presented here. Baseline results are computed using only utterance level normalization [2].

Table 2. EER (%) for normal speed test in NTT database. Utterance level normalization background speaker set - Top 10.

Model type	Feature	Background speakers			WMR	Base line
		All	Top10	Coh.		
4 m. full	cep	2.31	2.30	2.14	1.31	2.50
	c+Δc	1.51	1.48	1.33	0.84	1.64
8 m. full	cep	1.43	1.44	1.38	0.66	1.66
	c+Δc	1.09	1.09	0.96	0.52	1.18
32 m. diag.	cep	1.48	1.48	1.29	0.91	1.65
	c+Δc	1.14	1.13	1.00	0.95	1.29
64 m. diag.	cep	1.24	1.24	1.20	0.72	1.60
	c+Δc	0.87	0.88	0.86	0.60	1.07

Table 3. EER (%) for slow speed test in NTT database. Utterance level normalization background speaker set - Top 10.

Model type	Feature	Background speakers			WMR	Base line
		All	Top10	Coh.		
4 m. full	cep	3.36	3.33	3.00	1.94	3.79
	c+Δc	2.79	2.77	2.27	2.06	2.96
8 m. full	cep	2.18	2.16	2.06	1.45	2.46
	c+Δc	1.95	1.95	1.77	1.36	2.06
32 m. diag.	cep	2.60	2.62	2.16	1.50	2.90
	c+Δc	2.25	2.26	1.92	1.76	2.36
64 m. diag.	cep	2.76	2.76	2.23	1.57	3.15
	c+Δc	2.38	2.39	1.94	1.43	2.57

Table 4. EER (%) for fast speed test in NTT database. Utterance level normalization background speaker set - Top 10.

Model type	Feature	Background speakers			WMR	Base line
		All	Top10	Coh.		
4 m.	cep	2.78	2.75	2.65	1.92	3.07
full	c+Δc	2.15	2.15	1.93	1.29	2.26
8 m.	cep	1.89	1.88	1.66	1.11	2.01
full	c+Δc	1.27	1.26	1.09	0.80	1.43
32 m.	cep	2.91	2.91	2.51	1.90	3.06
diag.	c+Δc	2.71	2.71	2.58	1.79	2.88
64 m.	cep	2.42	2.44	2.06	1.48	2.65
diag.	c+Δc	2.44	2.43	2.44	1.28	2.66

"Model type" column of the tables specifies the GMM used in the experiments. "4 m. full" means GMM with 4 mixture densities with full covariance matrix and "32 m. diag" means GMM with 32 mixture densities with diagonal covariance matrix. The EER obtained using only mel-cepstral feature vectors ("cep") and both mel-cepstral and regressive feature vectors ("c+Δc") are presented in separate rows.

These results clearly show that the combination of frame and utterance level likelihood normalization outperforms the baseline with only utterance level normalization. All types of background speaker sets give better ERR. Among them the **Cohort** type is the best, while **All others** and **Top M** give almost the same results. **WMR** normalization, however, gives the lowest EER which in some cases is more than twice lower than the baseline results. Comparable results to our baseline system using the same database were also reported in [2]. When test utterances are uttered at slow or fast speed, the system performance degrades significantly, but this is due to the training of speaker models only on normal speed utterances. However, in both cases of slow and fast speed tests, all of our frame normalization techniques outperform the baseline system.

Table 5 shows the results for TIMIT database. In these experiments we used 168 speakers (the same as in [3]) with two test sentences per speaker and therefore rotating over all these speakers we had $168 \times 2 = 336$ true tests and $167 \times 2 \times 168 = 56112$ impostor tests available for experiments.

Table 5. EER (%) for TIMIT database. Utterance level normalization background speaker set - Top 20.

Model type	Feature	Background speakers			WMR	Base line
		All	Top20	Coh.		
4 m.	cep	0.71	0.71	0.67	0.48	0.72
full	c+Δc	0.60	0.60	0.56	0.39	0.61
8 m.	cep	0.42	0.42	0.38	0.16	0.43
full	c+Δc	0.45	0.45	0.41	0.15	0.46
16 m.	cep	0.39	0.39	0.35	0.16	0.40
full	c+Δc	0.45	0.45	0.40	0.19	0.46
32 m.	cep	0.58	0.58	0.51	0.09	0.59
diag.	c+Δc	0.65	0.65	0.58	0.13	0.66
64 m.	cep	0.57	0.57	0.53	0.24	0.57
diag.	c+Δc	0.76	0.76	0.70	0.19	0.76

In TIMIT database, the test and train conditions are the same which is very simple for the task and, consequently, it is more difficult to outperform the baseline performance. This is evident from the results of **All others** and **Top M** background speaker sets which in contrast to the multisession NTT database are the same as the baseline. The **Cohort** is slightly better, and **WMR** significantly reduces the EER to 0.09%. Using only utterance level likelihood normalization and different front-end analysis (16kHz f_s, 30 MFCC) Reynolds reported best EER of 0.24% in [3].

7. CONCLUSION

We have introduced and experimented with frame level likelihood normalization for speaker verification. A new technique, *Weighting Model Rank*, was also experimented. In combination with utterance level likelihood normalization both approaches showed better results in the speaker verification task compared to the standard accumulated likelihood method on both the TIMIT and NTT databases. As results for the NTT database show our frame level likelihood normalization is robust against both the variations of the speaker voices and speaking speeds. The WMR technique performed best on both databases reducing the equal error rate up to 0.52% for NTT and 0.09% for TIMIT database.

Our previous studies also showed that frame level likelihood normalization is effective for the speaker identification task and that, in general, any non-linear transformation (normalization) of the likelihoods at the frame level influences the speaker recognition performance.

REFERENCES

[1] A. Higgins, L. Bahler and J. Porter, "Speaker verification using randomized phrase prompting", Digital Signal Processing, Vol. 1, pp. 89-106, 1991.

[2] 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.

[3] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, Vol. 17, No. 1-2, pp.91-108, 1995.

[4] K. Markov and S. Nakagawa, "Frame level likelihood normalization for text-independent speaker identification using Gaussian mixture models", Proc. ICSLP, pp.1764-1767, 1996.

[5] D.A. Reynolds, "Large population speaker identification using clean and telephone speech", IEEE signal proc. letters, Vol.2, No.3, pp.46-48, March 1995.

[6] R.Duda and P.Hart, "Pattern Classification and Scene Analysis", John Wiley & Sons, 1973.

[7] 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.

[8] K. Fukunaga, "Introduction to statistical pattern recognition", Academic Press Inc., 1990.

[9] 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.