ROBUST MODEL FOR SPEAKER VERIFICATION AGAINST SESSION-DEPENDENT UTTERANCE VARIATION

Tomoko Matsui and Kiyoaki Aikawa

NTT Human Interface Laboratories
1-1, Hikari-no-oka, Yokosuka-shi, Kanagawa 239, Japan
tomoko@nttspch.hil.ntt.co.jp, aik@nttspch.hil.ntt.co.jp

ABSTRACT

This paper investigates a new method for creating speaker models robust against utterance variation in continuous distribution hidden-Markov-model-based speaker verification. In this method, the distribution of the session-independent features for each speaker is estimated by separately modeling the session-to-session utterance variation as two distinct variations: one session-dependent and the other session-independent. In practice, joint normalization of the session-dependent utterance variation and estimation of the parameters of speaker models is performed based on a speaker adaptive training algorithm. The resulting speaker models more accurately represent session-independent speaker characteristics, and the discriminatory capabilities of these models increases. In text-independent speaker verification experiments using data uttered by 20 speakers in 7 sessions over 16 months, we show that the proposed method achieves a 15% reduction in the error rate.

1. INTRODUCTION

Speaker verification systems are supposed to be used to judge the identities of individual speakers many times over a long period. In practical systems, a serious problem is the burden placed on each speaker to produce speech data, and as a result, speech data is often collected over several sessions and the amount of data collected in each session is usually small. In one session, a series of utterances is continuously recorded within a limited time, i.e., dozens of minutes. Although initial speaker models are created using such a small amount of data, the models are not robust against session-to-session utterance variation. The reason why this is such a difficult problem lies in the fact that the utterance variation is session-dependent and irregular. Generally, a large amount of data for each speaker would be saved over multiple sessions and the speaker model recreated using this large data set containing utterance variation. However, the spectral distributions of the large data set often exhibit a high degree of variance and the model represents fuzzy speaker characteristics. This may reduce the discriminatory capabilities of speaker models. Setlur et al. [1] reported a method of recreating the speaker model allowing for increased model complexity to better capture utterance variation. While such methods would be effective, it would be quite difficult to eliminate the effects of irregular utterance variation and determine the proper model complexity.

This paper investigates a method of creating speaker models robust against utterance variation that represents session-independent speaker characteristics more accurately by using the fixed-model complexity in continuous distribution hidden Markov model (HMM)-based speaker verification. Here, we assumed that session-to-session utterance-variation comprises two distinct variations: one being session-dependent caused by voice changes with time and by the difference in texts among sessions especially in text-independent systems, and the other is session-independent factors. Conceptually, this method attempts to remove session-dependent utterance variation rather than capture it. In this method, session-to-session utterance variation is modeled, and joint normalization of session-dependent utterance variation and the speaker model parameters are estimated based on the speaker adaptive training (SAT) algorithm [2][3][4]. This algorithm was originally developed for speaker adaptation in speech recognition and its development was motivated by the fact that variability in speaker-independent (SI) phoneme models is attributed to both phonetic variation and inter-speaker variation. The algorithm performs joint normalization of inter-speaker variation and estimation of the parameters of the SI models, leading to true SI models with reduced cross-unit overlap. The algorithm can be applied to our problem by replacing SI phoneme models and inter-speaker variation with speaker models and session-dependent utterance variation. We term the resulting speaker models as compact speaker models following the manner of [2].

In speaker verification, the likelihood normalization technique is essential in order to set stable thresholds for verification decision since the likelihood value covers a wide range of different texts spoken at different times, even by the same speaker. We previously reported the likelihood normalization method based on a posteriori probability [5]. In the normalization method, a likelihood value of a single phoneme- and speaker-independent pooled model, which is formed by pooling the features of all registered speakers, is used to normalize likelihood values of speaker models. This paper also investigates a method of creating a compact pooled model for compact speaker models.

2. COMPACT SPEAKER MODEL CREATION

In general, the speech data uttered by a speaker is assumed to be the sample that is drawn from a probability density function. Here, the speech data set uttered by a speaker in different sessions is assumed to be the sample set with different probability density functions corresponding to each session but have common session-independent speaker characteristics. According to this assumption, in our method, session-to-session utterance variation is modeled as a pair of distinct variations, one being session-dependent and the other session-independent. In the formulation, the HMM parameter set $\tilde{\theta}_s$ of speaker s estimated from the data of speaker s including only session-independent variation is mapped into HMM parameter set $\theta_s^{(t)}$ estimated from the data of speaker s and session t also including session-dependent variation by the model transformation function $G_s^{(t)}$ as follows:

$$\theta_s^{(t)} = G_s^{(t)}(\tilde{\theta}_s) \tag{1}$$

In this paper, through preliminary investigation of the model transformation functions, we consider $G_s^{(t)}$ of the form for mean vector μ_{sjk} of mixture component k in state j as

$$G_s^{(t)}(\tilde{\mu}_{sjk}) = \mu_{sjk}^{(t)} = \tilde{\mu}_{sjk} + b_s^{(t)}. \tag{2}$$

The optimum set of HMM parameter set $\tilde{\theta}_s$ for speaker s and the set of the model transformation functions of each session $\tilde{\mathcal{G}}_s = (\tilde{G}_s^{(1)}, \tilde{G}_s^{(2)}, \ldots, \tilde{G}_s^{(T)})$ are jointly estimated so as to maximize the likelihood using the SAT algorithm [2][3][4], i.e.,

$$(\tilde{\theta}_s, \tilde{\mathcal{G}}_s) = \arg \max_{(\theta_s, \mathcal{G}_s)} \prod_{t=1}^{T} \mathcal{L}(O_s^{(t)}; G_s^{(t)}, \theta_s) \tag{3}$$

where $O_s^{(t)}$ is the sample of speaker s and session t and $\mathcal{L}()$ is the HMM likelihood function.

The SAT algorithm is a 3-step optimization of the model transformation functions, mean and variance vectors (diagonal covariance HMMs). First, $\tilde{b}_s^{(t)}$ in the model transformation function of Eq. (2) is optimized according to the stochastic matching algorithm [6] as shown below.

$$\tilde{b}_{sl}^{(t)} = \frac{\sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{n=1}^{N_t} \gamma_{sjk}^{(t)}(n) \frac{o_{sl}^{(t)}(n) - \tilde{\mu}_{sjkl}}{\sigma_{sjkl}}}{\sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{n=1}^{N_t} \frac{\gamma_{sjk}^{(t)}(n)}{\sigma_{sjkl}}} \tag{4}$$

where subscript l denotes the lth component of the vectors, N_t the sample size of session t, $\gamma_{sjk}^{(t)}(n)$ denotes the probability of being in state j with mixture component k at time n given that the HMM of speaker s generates the observation vector $o_s^{(t)}(n)$, and σ_{sjk} denotes the variance vector.

Then mean vector $\tilde{\mu}_{sjk}$, variance vector $\tilde{\sigma}_{sjk}$ is optimized respectively, i.e.,

$$\tilde{\mu}_{sjk} = \frac{\sum_{t=1}^{T} \sum_{n=1}^{N_t} \gamma_{sjk}^{(t)}(n)(o_s^{(t)}(n) - \tilde{b}_s^{(t)})}{\sum_{t=1}^{T} \sum_{n=1}^{N_t} \gamma_{sjk}^{(t)}(n)} \tag{5}$$

$$\tilde{\sigma}_{sjkl} = \frac{\sum_{t=1}^{T} \sum_{n=1}^{N_t} \gamma_{sjk}^{(t)}(n)(o_{sl}^{(t)}(n) - \tilde{\mu}_{sjkl}^{(t)})^2}{\sum_{t=1}^{T} \sum_{n=1}^{N_t} \gamma_{sjk}^{(t)}(n)} \tag{6}$$

where $\tilde{\mu}_{sjk}^{(t)}(= \mu_{sjk} + \tilde{b}_s^{(t)})$ denotes the mean vector adapted to the sample of session t.

3. COMPACT POOLED MODEL CREATION

In the likelihood normalization method based on a posteriori probability [5], the a posteriori probability is used for verification decision, i.e.,

$$p(s_c|o) = \frac{p(o|s_c) \times p(s_c)}{\sum_i \{p(o|s_i) \times p(s_i)\}} \approx \frac{p(o|s_c)}{\sum_i p(o|s_i)} \tag{7}$$

where s_i denotes a speaker, s_c denotes the claimed speaker, and o denotes the input speech. The $p(s_i)$ is the probability for speaker i and is assumed to be constant for all speakers. The $p(o|s_c)$ is the probability of the claimed speaker's HMM. Then $\sum_i p(o|s_i)$ is approximated by a likelihood value for a pooled HMM made using the data set uttered by all registered speakers based on the maximum likelihood (ML) estimation since the number of calculations for the summation is enormous.

According to Eq. (7) , when using compact speaker HMMs, $\sum_i p(o|s_i)$ is the summation for the probabilities of compact speaker HMMs and should be approximated by a likelihood value for a compact pooled HMM made using the same data set as that for forming compact speaker HMMs, that is, the data set without session-dependent variation.

The parameters of a compact pooled HMM are estimated using $\tilde{b}_s^{(t)}$ optimized for each speaker and each session in Eq. (4) . Mean vector $\tilde{\mu}_{pjk}$, variance vector $\tilde{\sigma}_{pjk}$ of compact pooled model p is given by

$$\tilde{\mu}_{pjk} = \frac{\sum_{s=1}^{S} \sum_{t=1}^{T_s} \sum_{n=1}^{N_{st}} \gamma_{pjk}^{(t)}(n)(o_s^{(t)}(n) - \tilde{b}_s^{(t)})}{\sum_{s=1}^{S} \sum_{t=1}^{T_s} \sum_{n=1}^{N_{st}} \gamma_{pjk}^{(t)}(n)} \tag{8}$$

$$\tilde{\sigma}_{pjkl} = \frac{\sum_{s=1}^{S} \sum_{t=1}^{T_s} \sum_{n=1}^{N_{st}} \gamma_{pjk}^{(t)}(n)(o_{sl}^{(t)}(n) - \tilde{\mu}_{pjkl}^{(t)})^2}{\sum_{s=1}^{S} \sum_{t=1}^{T_s} \sum_{n=1}^{N_{st}} \gamma_{pjk}^{(t)}(n)} \tag{9}$$

where $\tilde{\mu}_{pjk}^{(t)}(= \mu_{pjk} + \tilde{b}_s^{(t)})$ denotes the mean vector mapped by the model transformation function of speaker s and session t.

4. EXPERIMENTAL CONDITIONS

The proposed method was evaluated in text-independent speaker verification experiments. The database comprises sentence data uttered by 20 male speakers; 10 speakers were used as customers and the remainder were used as impostors. The sentences were selected from phonetically balanced sentences [7] and were read. The speech was recorded in seven sessions (T1-7) over 16 months and was recorded in the same recording room using the same microphone for all speakers and for all sessions. The sampling rate was 12 kHz. The cepstral coefficients were calculated by LPC analysis with an order of 16, a frame period of 8 ms, and a frame length of 32 ms. We used 1-state, 16-Gaussian-mixture, diagonal covariance HMMs as speaker models and a 1-state, 64-Gaussian-mixture, diagonal covariance HMM as a pooled model. For training, initial speaker models

Case	X	A	B	C	D	E
Training	T1 [5]	T1,T2 [10]	T1-3 [15]	T1-4 [20]	T1-5 [25]	T1-6 [30]
Testing	T2	T3	T4	T5	T6	T7

Table 1: Sessions of sentences for training and testing ([ ]: total number of training sentences!K

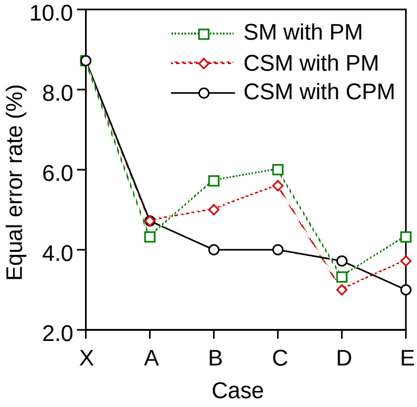

Figure 1: Equal error rate (%) comparison of several methods.

were created using five sentences from session T1 and the models were recreated also using five sentences from the next session respectively. The texts were varied from customer to customer and from session to session. The average duration of each sentence was 4.2 sec. For testing, the beginning 1 sec. of each of three sentences from the subsequent session for training was evaluated individually. The sentences for testing were different from those for training and were the same for all customers and impostors and all recording sessions. Table 3 lists sessions (case X, A-E) of sentences for training and testing. In the experiments, the likelihood normalization method based on a posteriori probability was used. The threshold was set a posteriori for individual speakers to equalize the probability of false acceptance and false rejection, and an equal error rate was used for evaluation.

5. RESULTS

Figure 1 shows the equal error rates for several combinations of speaker models (SMs), compact speaker models (CSMs), a pooled model (PM), and a compact pooled model (CPM) for each case. In the "SM with PM" method, for instance, each speaker model was conventionally recreated based on the ML estimation using all available data of the speaker in the case. Likelihood values of speaker models were normalized using a likelihood value of a pooled model recreated based on the ML estimation using all available data of all registered speakers in the case. The "CSM with CPM" method performed stably for each case and was the best on average. The error reduction rate compared with the "SM and PM" method was 15% on average.

16-SM with PM	20-SM with PM	24-SM with PM	16-CSM with CPM
4.7	4.7	4.3	4.0

Table 2: Equal error rates (%) averaged over cases A to E with different numbers of mixture components.

Since compact speaker models are assumed to be created using data including only session-independent utterance variation, the models are expected to achieve the same performance with fewer parameters than those needed for conventional models created using data also including session-dependent utterance variation. Table 2 lists the equal error rates averaged over cases A to E for the "SM with PM" method using speaker models with different numbers of mixture components. The "16-CSM with CPM" method performed better than the "24-SM with PM" method in which speaker models were represented using 24-Gaussian-mixture HMMs. These results indicate that compact speaker models efficiently represent session-independent speaker characteristics with the fewer parameters.

Cepstrum mean normalization (CMN) is a well-known technique for canceling the effects of channels and utterance variation in speaker recognition [8][9]. CMN has the advantage of normalizing session-dependent utterance variation, but it has the disadvantage of also normalizing statistical speaker characteristics on the averaged cepstrum for each utterance which is effective in speaker recognition [10]. Table 3 compares the equal error rates with/without CMN. The "1 session" method uses speaker models and a pooled model made only using data at the latest session in the case. The "CSM with CPM" method without CMN performed best. While the rates for the "1 session" and "SM with PM" methods without CMN were higher than those with

CMN	1 session	SM with PM	CSM with CPM
-	10.2	4.7	4.0
√	7.8	4.5	4.6

Table 3: Equal error rates (%) averaged over cases A to E for several methods with/without CMN.

Case	A	B	C	D	E
$\|$ SM $\|$ /	1.01	1.05	1.05	1.05	1.05
$\|$ CSM $\|$	[1.06]	[1.16]	[1.13]	[1.14]	[1.12]

Table 4: Averaged and [maximum] ratios of mixture-variance norm for SM and CSM.

CMN, the rates for the "CSM with CPM" method without CMN were lower than those with CMN. These results indicate that in "1 session" and "SM with PM" methods, the advantage of CMN overcomes its disadvantage and in the "CSM with CPM" method, since compact speaker models have effectively normalized the effects of session-dependent utterance variation, the advantage is nullified and the remaining disadvantage decreases the performance.

6. DISCUSSION

In order to confirm the assumption in Section 2 , variance vector norms of mixture-components for conventional SMs and CSM have been examined. Variance vector $\hat{\sigma}_{sjkl}$ for SMs is given by

$$\hat{\sigma}_{sjkl} = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N_t} \gamma'^{(t)}_{sjk}(n)(o^{(t)}_{sl}(n) - \hat{\mu}_{sjkl})^2}{\sum_{t=1}^{T}\sum_{n=1}^{N_t} \gamma'^{(t)}_{sjk}(n)} \qquad (10)$$

where $\gamma'^{(t)}_{sjk}(n)$ denotes the probability of being in state j with mixture component k at time n given that the HMM made based on the ML estimation using all available data generates observation vector $o^{(t)}_s(n)$. We can explain Eq. (6) and Eq. (10) as a probabilistic average of each variance vector for session 1 to T with a weight corresponding to each data length. When the assumption is true, each variance vector of each session for CSM is closer to the uniformly minimum variance unbiased estimator than that for SM. Therefore, it is expected that the variance vector norms of CSM should be smaller than those for SM.

Table 4 lists the averaged and maximum ratios of the variance vector norms for SM and CSM. The ratios were higher than 1.0 and hence the assumption would be reasonable.

7. CONCLUSION

We have presented a new method for creating compact speaker models using the SAT algorithm and a method for creating a compact pooled model for compact speaker models in the likelihood normalization method based on a posteriori probability. Text-independent speaker verification experiments showed that the combination of these methods was effective and robust against session-to-session utterance variation. Comparison of the performance between the proposed method and methods using conventional speaker models with a larger number of mixture-components showed that compact speaker models efficiently represent session-independent speaker characteristics with fewer parameters. Moreover, comparison of the performance between methods with/without CMN showed that the proposed method effectively normalized the effects of session-dependent utterance variation.

Further study includes investigation of more effective model transformation functions for session-to-session utterance variation and examination of the method using a larger number of speakers and data in real fields.

ACKNOWLEDGMENTS

We thank the members of the speech and acoustics laboratory of the NTT Human Interface Laboratories for their valuable and stimulating discussions.

8. REFERENCES

[1] A. Setlur and T. Jacobs, *Results of a speaker verification service trial using HMM models*, Proc. Eurospeech, pp. I-53-56, 1995.

[2] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, *A Compact Model for Speaker-Adaptive Training*, Proc. ICSLP, pp. 1137-1140, 1996.

[3] T. Anastasakos, J. McDonough, and J. Makhoul, *Speaker Adaptive Training: A Maximum Likelihood Approach to Speaker Normalization*, Proc. ICASSP, pp. 1043-1046, 1997.

[4] D. Pye and P.C. Woodland, *Experiments in Speaker Normalization and Adaptation for Large Vocabulary Speech Recognition*, Proc. ICASSP, pp. 1047-1050, 1997.

[5] T. Matsui and S. Furui, *Likelihood normalization using a phoneme- and speaker-independent model for speaker verification*, Speech Communication, Vol. 17, No. 1-2, pp. 109-116, 1995.

[6] A. Sankar and C.-H. Lee, *Robust speech recognition based on stochastic matching*, Proc. ICASSP, pp. I-121-124, 1995.

[7] H. Kuwabara, Y. Sagisaka, K. Takeda, and M. Abe, *Construction of ATR Japanese speech database as a research tool*, ATR Tech. Rep. TR-I-0086, 1989.

[8] B.S. Atal, *Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification*, J. Acoust. Soc. Amer., 55, 6, pp. 1304-1312, 1974.

[9] S. Furui, *Cepstral analysis technique for automatic speaker verification*, IEEE Trans. ASSP, 29, 2, pp. 254-272, 1981.

[10] S. Furui, F. Itakura and S. Saito, *Talker recognition by longtime averaged speech spectrum*, Trans. IECE, 55-A, 10, pp. 549-556, 1972.