ROBUST SPEAKER VERIFICATION INSENSITIVE TO SESSION-DEPENDENT UTTERANCE VARIATION AND HANDSET-DEPENDENT DISTORTION

Tomoko Matsui *Kiyoaki Aikawa*

NTT Human Interface Laboratories
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan

ABSTRACT

This paper investigates a method of creating robust speaker models that are not sensitive to session-dependent (SD) utterance-variation and handset-dependent (HD) distortion for hidden Markov model (HMM)-based speaker verification systems in a real telephone network. We recently reported a method of creating session-independent (SI) speaker-HMMs that are not sensitive to SD utterance-variation. In that method, the distortion function that transforms SI speaker-HMMs to SD speaker-HMMs is introduced, and the parameters in the function and the speaker-HMM parameters are jointly estimated using a speaker adaptive training algorithm. This paper proposes a new method that is less sensitive to SD utterance-variation and HD distortion than the previous method. This new idea focuses on different difficulties in estimating parameters in distortion functions for SD utterance-variation and HD distortion. In text-independent verification experiments with recorded data from different handsets, the error reduction rate of the improved method compared with that of the conventional method of cepstral mean normalization is 24% when each speaker-HMM is recreated using data uttered in five sessions.

1. INTRODUCTION

For speaker verification systems in a real telephone network, input speech is distorted not only by session-dependent (SD) utterance-variation caused by the changes in vocal-tract conditions but also by the distortion caused by the acoustic difference of handsets (including transmission line characteristics). This causes serious degradation of the system performance. In practical systems, the burden placed on each speaker to produce speech data is a problem that requires careful consideration. As a result, speech data is often collected using some fixed handsets in a house or office 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 seconds. Although initial speaker models are created using such a small amount of data recorded from some fixed handsets, the models are not robust against SD utterance-variation and handset-dependent (HD) distortion.

For SD utterance-variation, we recently reported a method for creating robust speaker models that represent session-independent (SI) speaker characteristics more accurately by using the fixed-model complexity in HMM-based speaker verification [1]. The reason why SD utterance-variation is a difficult problem lies in the fact that it is session-dependent and irregular. Generally, a large amount of data for each speaker would be saved over multiple sessions and a speaker model would be 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. In our previous method [1], it was assumed that session-to-session utterance-variation comprises two distinct variations: one being SD caused by voice changes with time and by the difference in texts among sessions especially in text-independent systems, and the other is SI factors. Conceptually, the proposed method attempts to remove SD utterance-variation. In this method, SD utterance-variation is modeled as a distortion function, which is an HMM parameter transformation function, and transforms the SI speaker-HMM to the SD speaker-HMM. The parameters in the function and the SI speaker-HMM parameters are jointly estimated using the speaker adaptive training (SAT) algorithm [2][3][4]. Text-independent speaker verification experiments with data uttered by 20 male speakers over multiple sessions (3-month intervals) that were recorded using the same condenser microphone showed that this method more effectively normalized the effects of SD utterance-variation than the conventional method of cepstral mean normalization (CMN) [5][6].

This paper reports an improvement on our previous method making it less sensitive to SD utterance-variation and HD distortion. Here, we assume that session-to-session variation in speech data comprises two distinct variations: one being SD including SD utterance-variation and HD distortion, and the other is SI. Basically, the improved method attempts to remove only SD utterance-variation. We consider that although it is difficult to estimate irregular SD utterance-variation from a small amount of input speech, HD distortion can be represented in a simpler form and estimated from the input speech. In the improved method, the SI speaker-HMM is first estimated using the composite function of the distortion functions for SD utterance-variation and HD distortion based on the SAT algorithm. Then, the parameter in the distortion function of the HD distortion is estimated from input speech as a common bias for each mixture-mean vector, and the SI speaker-HMM is transformed to the SI-but-HD speaker-HMM by using the distortion function. This SI-but-HD speaker-HMM is used to judge the identities of individual speakers. In the improved method, the distortion function of SD utterance-variation is formulated as a linear transformation function with scale and bias factors for each mixture-mean vector in the speaker-HMM, and the performance levels of the functions with/without a scale factor are compared.

2. ROBUST MODEL CREATION

In the proposed method, for training, the SI speaker-HMM is estimated, and for testing, the distortion function of the HD distortion in input speech is estimated. The SI-but-HD speaker-HMM is created by transforming the SI speaker-HMM using the distortion function. The following sections explain how we create the SI speaker-HMM and estimate the distortion function of the HD distortion in input speech.

2.1. Creating session-independent speaker model

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

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

In a similar way, the HMM parameter set $\tilde{\theta}_s$ is mapped into HMM parameter set $\theta_s^{(t)}$ estimated from the data of speaker s and session t distorted by HD distortion modeled as the distortion function $H_s^{(t)}$ as follows:

$$\theta_s^{(t)} = H_s^{(t)}(\tilde{\theta}_s). \tag{2}$$

In this paper, $G_s^{(t)}$ and $H_s^{(t)}$ are defined as the forms for mean vector μ_{sjk} of mixture component k in state j as

$$G_s^{(t)}(\tilde{\mu}_{sjk}) = \mu_{sjk}^{(t)} = a_s^{(t)} \cdot \tilde{\mu}_{sjk} + b_s^{(t)}, \tag{3}$$

$$H_s^{(t)}(\tilde{\mu}_{sjk}) = \mu_{sjk}^{(t)} = \tilde{\mu}_{sjk} + h_s^{(t)}. \tag{4}$$

Here, $a_s^{(t)}$ is a scale vector for each mean vector, and $b_s^{(t)}$ and $h_s^{(t)}$ are bias vectors. Then, composite function $F_s^{(t)}$ of $G_s^{(t)}$ and $H_s^{(t)}$ is defined as follows:

$$F_s^{(t)}(\tilde{\mu}_{sjk}) = \mu_{sjk}^{(t)} = a_s^{(t)} \cdot \tilde{\mu}_{sjk} + z_s^{(t)} \tag{5}$$

where $z_s^{(t)} = b_s^{(t)} + h_s^{(t)}$.

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

$$(\tilde{\theta}_s, \tilde{\mathcal{F}}_s) = \arg \max_{(\theta_s, \mathcal{F}_s)} \prod_{t=1}^{T} \mathcal{L}(O_s^{(t)}; F_s^{(t)}, \theta_s) \tag{6}$$

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 distortion functions, mean, and variance vectors (diagonal covariance HMMs). Terms $\tilde{a}_s^{(t)}$ and $\tilde{z}_s^{(t)}$ in the composite function of Equation (5) are optimized according to the maximum likelihood linear regression algorithm [7].

2.2. Handset-dependent distortion estimation

Since the distortion function of HD distortion in input speech $H_s^{(u)}$ (u denotes the session in which the speech was uttered) is speaker-independent, we assumed that it can be approximately estimated so as to maximize the likelihood of all speaker-HMMs $\{\theta_1, \theta_2, \ldots, \theta_s, \ldots, \theta_r, \ldots, \theta_R\}$ for input speech, i.e.,

$$\tilde{H}_s^{(u)} = \arg \max_{H_s^{(u)}} \prod_{r=1}^{R} \mathcal{L}(O_s^{(u)}; H_r^{(u)}, \theta_r) \tag{7}$$

where s denotes the claimed speaker and r denotes a speaker.

Here, since the amount of calculation of Equation (7) is linearly proportional to the population of the registered speakers, the product is approximated using pooled HMM θ_p made using the data set uttered by all registered speakers based on the maximum likelihood (ML) estimation as follows:

$$\tilde{H}_s^{(u)} = \arg \max_{H_s^{(u)}} [\mathcal{L}(O_s^{(u)}; H_s^{(u)}, \theta_s) \cdot \mathcal{L}(O_s^{(u)}; H_s^{(u)}, \theta_p)]. \tag{8}$$

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

Table 1: Sessions of Sentences for Training and Testing ([ ]: Total number of training sentences!K

3. 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 remaining speakers were used as impostors. The sentences were selected from phonetically balanced sentences [9] and were read. Originally, the speech was recorded in six sessions (T1-6) over 15 months and was recorded in the same recording room using the same microphone for all speakers and for all sessions. Here, the speech was re-recorded as telephone speech randomly using five kinds of handsets. The frequency range was 300-3400 Hz. 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 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 1 lists sessions (case X, A-D) of sentences for training and testing. In the experiments, the likelihood normalization method based on a posteriori probability was used [1][8]. 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.

4. RESULTS

Case	X	A	B	C	D
ML	30.7	24.7	24.3	20.0	14.5
SI(bias only)	30.7	26.7	24.3	21.7	14.0
SI(bias+scale)	30.7	25.7	23.0	19.7	14.0
SI(bias only)+HD	24.1	21.0	19.7	19.7	11.0
SI(bias+scale)+HD	24.1	20.3	19.7	18.7	12.3

Table 2: Equal Error Rate (%) Comparison of ML and Combinations of SI and HD Methods

Table 2 lists the equal error rates for ML and combinations of SI and HD methods for each case. In the "ML" method, 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 [8]. In the "SI(bias only)" and "SI(bias+scale)" methods, SI speaker-HMMs were estimated using a linear transformation function with/without the scale factor as the distortion function, and were used to judge the identities of individual speakers. The "SI(bias only)+HD" and "SI(bias+scale)+HD" methods represent our proposed method and use SI-but-HD speaker-HMMs. Likelihood values of SI and SI-but-HD speaker-HMMs were normalized using a likelihood value of an SI and SI-but-HD pooled model estimated using the same distortion function as that for SI and SI-but-HD speaker-HMMs [1].

The "SI+HD" method performed stably for each case, and the "SI(bias only)+HD" method showed the best performance in case D. The error reduction rate compared with the "ML" method was 24% in case D and 15% on average. The "SI(bias+scale)" method performed slightly better than the "SI(bias only)" method, and the performance levels of the "SI(bias only)+HD" and "SI(bias+scale)+HD" methods were almost the same. The averaged value of $a_s^{(t)}$ in Equation (5) for each session and each vector order was 0.98. This indicates that when using 1-state, 16-Gaussian-mixture HMMs as speaker models, the stretch of the mean vector space in the speaker-HMM varies only slightly among sessions, and incorporating the scale factor in the distortion function is not effective. The speech data used in these experiments were originally uttered at some fixed speaking rate for all sessions. It can be considered that this controlled speaking rate makes the stretch of the mean vector space unvarying.

5. DISCUSSION

Cepstrum mean normalization (CMN) is a well-known technique for canceling the effects of channel and utterance variation in speaker recognition [5][6].

Figure 1 compares the equal error rates for the "ML+CMN" method and the "SI+HD" method. In the "ML+CMN" method, speaker models were made using the "ML" method with CMN. When only speech data uttered over less than four sessions are available (cases A

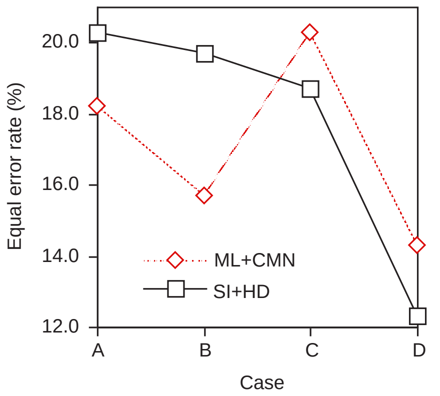

Figure 1: Equal error rates (%) for ML+CMN and SI+HD methods.

and B), the "ML+CMN" method performed better than the "SI+HD" method. On the other hand, when speech data uttered over four or more sessions are available (cases C and D), the "SI+HD" method performed better than the "ML+CMN" method. CMN has the advantage of normalizing SD variation, but it has the disadvantage of also normalizing statistical speaker-dependent features included in the long-term mean cepstrum for each utterance, which is effective in speaker recognition [10]. It can be considered that even when the number of sessions for speech data is increased, statistical SI speaker characteristics cannot be represented well in speaker-HMMs with CMN because of the disadvantage of CMN, and the performance may not necessarily improve.

6. CONCLUSION

We reported a method of creating SI-but-HD speaker models that are less sensitive to SD utterance-variation and HD distortion for speaker verification over the public switched telephone network. Text-independent speaker verification experiments showed that the proposed method was effective and robust against session-to-session variation in input speech. For the distortion function of SD utterance-variation, the performance levels of the linear transformation functions with only a bias factor and with both scale and bias factors were compared. It was shown that when using 1-state, 16-Gaussian-mixture HMMs as speaker models, the stretch of the mean vector space in the speaker-HMM varies only slightly for each session, and incorporating the scale factor in the distortion function is not effective. Moreover, a performance comparison of the conventional ML method with CMN and the proposed method showed that when speech data uttered over four or more sessions are available, the proposed method performed better than the conventional ML method with CMN because statistical SI speaker characteristics cannot be represented well in speaker-HMMs with CMN.

ACKNOWLEDGMENTS

We thank the members of the Speech and Acoustics Laboratory of the NTT Human Interface Laboratories for their valuable and stimulating discussions.

REFERENCES

1. T. Matsui and K. Aikawa, *Robust model for speaker verification against session-dependent utterance variation*, Proc. ICASSP, pp. I-117-120, 1998.
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. 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.
6. S. Furui, *Cepstral analysis technique for automatic speaker verification*, IEEE Trans. ASSP, 29, 2, pp. 254-272, 1981.
7. C. J. Leggetter and P. C. Woodland, *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models*, Computer Speech and Language, Vol. 9, pp. 171-185, 1995.
8. 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.
9. 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.
10. S. Furui, F. Itakura, and S. Saito, *Talker recognition by longtime averaged speech spectrum*, Trans. IECE, 55-A, 10, pp. 549-556, 1972.