EVALUATION OF MODEL ADAPTATION BY HMM DECOMPOSITION ON TELEPHONE SPEECH RECOGNITION

Tetsuya Takiguchi[1] *Satoshi Nakamura*[1] *Kiyohiro Shikano*[1] *Masatoshi Morishima*[2] *Toshihiro Isobe*[2]

[1]Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN

[2]Laboratory for Information Technology, NTT DATA Corporation
66-2, Horikawa-cho, Saiwai-ku, Kawasaki-shi, Kanagawa, 210-0913, JAPAN

E-mail: tetuy-t@is.aist-nara.ac.jp

ABSTRACT

In this paper, we evaluate performance of model adaptation by the previously proposed HMM decomposition method[1] on telephone speech recognition. The HMM decomposition method separates a composed HMM into a known phoneme HMM and an unknown noise and channel HMM by maximum likelihood (ML) estimation of the HMM parameters. A transfer function (telephone channel) HMM is estimated using adaptation speech data by applying the HMM decomposition twice in the linear spectral domain for noise and in the cepstral domain for channel. The telephone speech data for evaluation are recorded through 10 kinds of ordinary analog telephone handsets and cordless telephone handsets. The test results show that the average phrase accuracy with the clean speech HMMs is 60.9% for the ordinary analog telephone handsets, and 19.6% for the cordless telephone handsets. By the HMM decomposition method, the average phrase accuracy is improved to 78.1% for the ordinary analog telephone handsets, and 50.5% for the cordless telephone handsets.

1. INTRODUCTION

Many methods have been proposed to cope with problems caused by additive noise and convolutional distortion in robust speech recognition. Speech enhancement and model compensation approaches are two common examples among them. For the speech enhancement approach, spectral subtraction for additive noise and cepstral mean normalization(or signal bias removal) for convolutional distortion have been proposed (e.g., [2, 3, 4, 5, 6]). For the model compensation approach, conventional multi-template technique, model adaptation (e.g., [12, 13]) as well as model (de-)composition methods (e.g., [7, 8, 9, 10, 11, 14, 15]) have been developed.

In our previous paper [1], we proposed an HMM decomposition method. The HMM decomposition method deals with the model parameter instead of the series of the observed signal, and estimates the HMM parameters based on maximum likelihood (ML) estimation. Its effectiveness is confirmed by the word recognition experiments on the real distant-talking speech.

This paper reports the performance of the HMM decomposition method on the telephone speech recognition. The telephone speech data for evaluation are recorded using 10 kinds of the ordinary analog telephone handsets and the cordless telephone handsets in a soundproof room, through the public telephone network as shown in Figure 1.

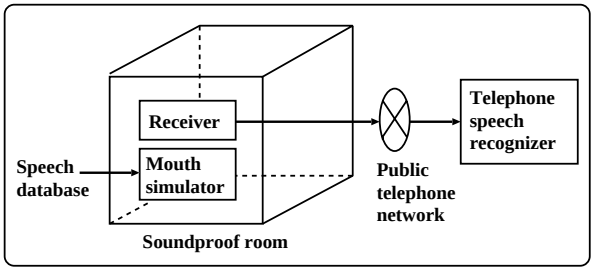

Figure 1: Recording condition of telephone speech

2. TELEPHONE SPEECH DATA

Figure 1 shows the recording condition of the telephone speech. Utterances from 60 speakers in the ASJ(Acoustical Society of Japan) continuous speech database are outputted through a mouth simulator, and inputted into 10 kinds of the ordinary analog telephone handsets and the cordless telephone handsets in the soundproof room. Then, their speech are recorded through the public telephone network. Ten kinds of telephone handsets are CANON (CF-H1CL), KENWOOD (IS-W757), NEC (Speax23 CL), NTT (CP-D40), PANASONIC (VE-D67L-K), PIONEER (TF-JP50), SANYO (TEL-L710), SHARP (CJ-H7-B), SONY (SPP-A600) and VICTOR (TN-DJ1-B). Each telephone handset consists of an ordinary analog telephone handset and a cordless telephone handset.

Figure 2 shows the log power spectrum of the clean speech and the telephone speech, which are digitized at an 8kHz sampling rate. In the case of the speech through the cordless telephone handset, the shape over 3kHz is distorted. The SNRs of the ordinary analog telephone handsets and the cordless telephone handsets are 25.1dB and 20.3 dB, respectively. Their SNRs are calculated by

$$SNR \sim 10\log_{10} \frac{\frac{1}{l}\sum_{t=1}^{l} o(t)^2}{\frac{1}{m}\sum_{t=1}^{m} n(t)^2},$$

where $o(t)$ and $n(t)$ denote the observed speech and the noise at time t, respectively. l and m are the number of total frames of speech data and the number of total frames of noise data, respectively.

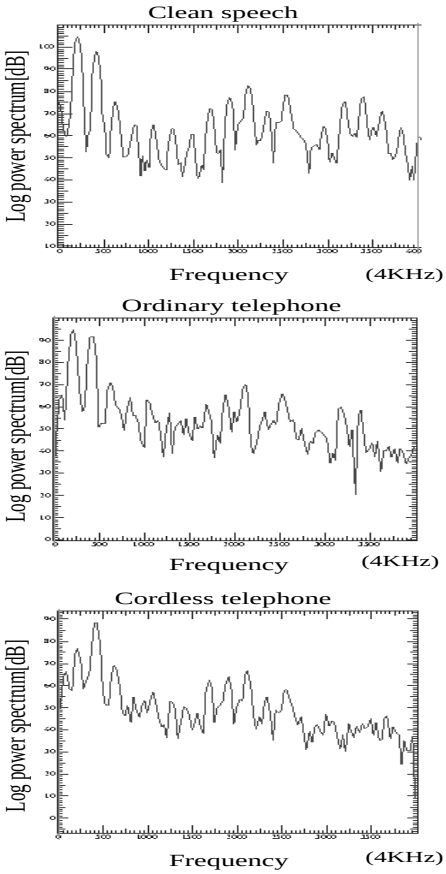

Figure 2: Log power spectrum /u/

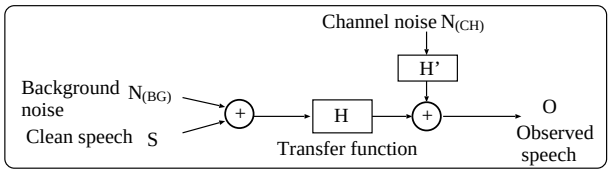

Figure 3: Environment model for telephone speech

3. HMM DECOMPOSITION

The HMM decomposition method separates a composed HMM into a known phoneme HMM and an unknown noise and channel HMM by maximum likelihood (ML) estimation of the HMM parameters[1].

Figure 3 shows an environment model for the telephone speech. The observed speech $O(\omega;m)$ is represented by

$$\begin{aligned} O(\omega;m) &= \{S(\omega;m)+N_{(BG)}(\omega;m)\}\cdot H(\omega;m) \\ &\quad +N_{(CH)}(\omega;m)\cdot H'(\omega;m) \\ &= S(\omega;m)\cdot H(\omega;m)+N(\omega;m), \end{aligned}$$

where

$$N(\omega;m)=N_{(BG)}(\omega;m)\cdot H(\omega;m)+N_{(CH)}(\omega;m)\cdot H'(\omega;m).$$

$S(\omega;m)$, $N_{(BG)}(\omega;m)$, $N_{(CH)}(\omega;m)$, and $N(\omega;m)$ denote the clean speech, the background noise, the channel noise and the observed noise at frame m and frequency ω, respectively. $H(\omega;m)$ and $H'(\omega;m)$ are transfer function. Accordingly, a composed HMM of the observed speech in the linear spectral domain is represented by

$$\lambda_{SH+N}=\mathrm{Exp}\{\ \mathrm{Cos}\ (\lambda_{S_{cep}}\oplus\lambda_{H_{cep}})\}\oplus\lambda_{N_{lin}}, \qquad (1)$$

where λ and $\oplus$ denote a set of model parameters and a model composition procedure, respectively. Exp and Cos are exponential transform of the distribution function and cosine transform of the distribution function, respectively. According to the equation (1), the estimation equation of the transfer function HMM is written as follows in the cepstral domain:

$$\lambda_{H_{cep}}=\mathrm{Cos}^{-1}\{\mathrm{Log}(\ \lambda_{SH+N}\ominus\ \lambda_{N_{lin}})\}\ominus\lambda_{S_{cep}}, \quad (2)$$

where *cep* and *lin* denote the cepstral domain and the linear spectral domain, respectively. $\ominus$ denotes a model decomposition procedure. Cos^{-1} and Log are inverse cosine transform of the distribution function and logarithm transform of the distribution function, respectively. The equation (2) shows that the HMM decomposition method is applied twice in the linear spectral domain and in the cepstral domain, where the transfer function HMM is estimated in noisy environment. Firstly, the HMM decomposition method is applied in the linear spectral domain to estimate the telephone speech HMMs which are free from the influences of noises. The obtained telephone speech HMMs are converted to the cepstral domain. Then, the HMM decomposition method is applied again to estimate the transfer function HMM. The procedure is summarized in Figure 4.

4. EXPERIMENTS AND RESULTS

4.1. Experimental condition

All experiments in this paper are conducted on the telephone speech data which we described in the section 2. About 7500 sentences from 25 males and 25 females are used for the training. Five males and five females for the testing are not used in the training. Each testing speaker utters only one sentence for adaptation for each handset.

We choose 55 context independent phonemes as the clean speech units. Each phoneme is modeled by a single left-to-right 3-state tied-mixture HMM with 3 self-transition loops and without state skipping. Sixteen mel-frequency cepstral coefficients(MFCC) with their first order differentials (ΔMFCC), and the first order differentials for normalized logarithmic energy (Δpower) are calculated as the observation vector for each frame. There are total 256 Gaussian mixture components with diagonal covariance matrices shared by all of the models for MFCC and ΔMFCC, respectively. There are 64 Gaussian mixture components shared by all of the models for Δpower.

For environment adaptation, a single Gaussian is employed to model the noise and the transfer function. Only mean vector is estimated for the transfer function in this experiment.

The phrase recognition experiment is carried out using continuous sentence speech. Each sentence includes 6 $\sim$ 7

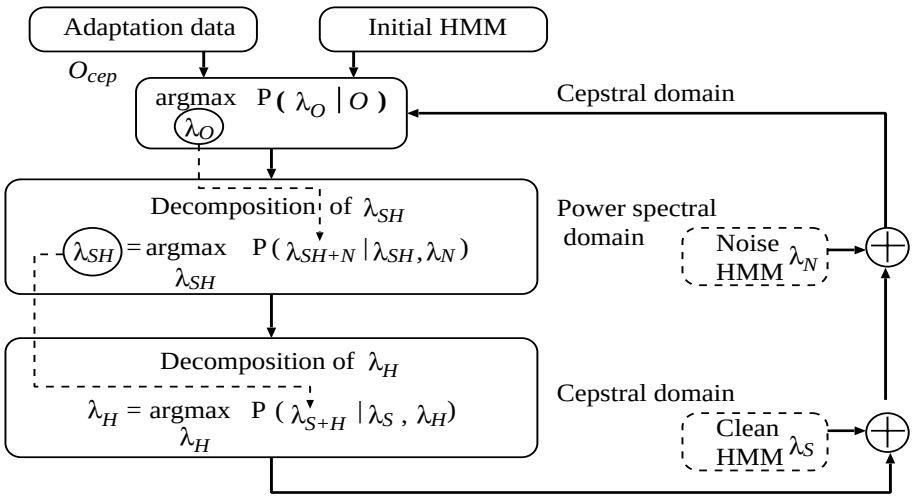

Figure 4: Parameter estimation by HMM decomposition

Table 1: Phrase accuracy[%] for ordinary analog telephone handsets

Model	HMM-S (Clean)	CMS	HMM -SH	HMM -SN	HMM -SHN	HMM-TELE (ordinary tele.)	HMM-TELE(ordinary and cordless)
Noise compensation	×	×	×	○	○	-	-
Channel compensation	×	○	○	×	○	-	-
Phrase accuracy	60.9	74.7	68.6	70.1	78.1	77.7	72.7

Table 2: Phrase accuracy[%] for cordless telephone handsets

Model	HMM-S (Clean)	CMS	HMM -SH	HMM -SN	HMM -SHN	HMM-TELE (cordless tele.)	HMM-TELE(ordinary and cordless)
Noise compensation	×	×	×	○	○	-	-
Channel compensation	×	○	○	×	○	-	-
Phrase accuracy	19.6	42.0	29.1	30.3	50.5	61.0	60.5

phrases on average. In this task, the ASJ database is divided into 10 subsets. Each subset consists of 50 sentences except one subset which consists of 53 sentences. One typical subset of this task is 323 phrases with a phrase perplexity of 323 on average. Each speaker utters 3 subsets through one telephone handset.

4.2. Experimental results

The points to be investigated are:

- improvement of recognition rate by the HMM composition and the HMM decomposition method,
- comparison with cepstral mean subtraction(CMS),

and

- comparison with matched condition.

Table 1 and Table 2 show the average phrase accuracy[%] for 10 kinds of the ordinary analog telephone handsets and the cordless telephone handsets, respectively. The phrase accuracy with the Clean HMMs(indicated as HMM-S) is 79.2% for the clean speech. The telephone speech, however, decreases the phrase accuracy to 60.9% for the ordinary analog telephone handsets, and 19.6% for the cordless telephone handsets.

The phrase accuracy with the HMM-SN, composed of the HMM-S and the noise HMM, is improved to 70.1% for the ordinary analog telephone handsets, and 30.3% for the cordless telephone handsets. By applying the HMM decomposition method twice in the linear spectral domain and in the cepstral domain, HMM-SHN, the phrase accuracy is improved from 60.9% to 78.1% for the ordinary analog telephone handsets, and from 19.6% to 50.5% for the cordless telephone handsets with one adaptation sentence.

Table 1 and Table 2 also include the average phrase accuracy for 10 kinds of the telephone handsets in the matched condition. The HMM phonemes, HMM-TELE(ordinary tele.) are trained by the speech data through 10 kinds of the ordinary analog telephone handsets. The HMM phonemes, HMM-TELE(cordless tele.), are trained by the speech data through 10 kinds of the cordless telephone handsets. The HMM phonemes, HMM-TELE(ordinary and cordless), are trained by the speech data through 10 kinds of the ordinary analog telephone handsets and the cordless telephone handsets. The phrase accuracy with the HMM-TELE(ordinary tele.) is 77.7% for the ordinary analog telephone handsets. The phrase accuracy with the HMM-TELE(cordless tele.) is 61.0% for the cordless telephone handsets. On the other hand, the phrase accuracy with the HMM-TELE(ordinary and cordless) is decreased to 72.7% for the ordinary analog telephone handsets, and

Table 3: Comparison with CMS(ordinary/cordless)

Estimation data	CMS	HMM-SH
adaptation 1	74.7% / 42.0%	-
adaptation 2	72.6% / 38.6%	68.6% / 29.1%

Table 4: Comparison with matched condition

Input	HMM-S	HMM -SHN	HMM-TELE (matched handset)
matched handset	64.5%	80.1%	86.6%

60.5% for the cordless telephone handsets. This is caused by the mismatched condition between the ordinary analog telephone handsets and the cordless telephone handsets.

Table 3 shows the comparison with CMS. When the HMM decomposition method is applied once in the cepstral domain(indicated as HMM-SH), the phrase accuracy is decreased to 68.6% for the ordinary analog telephone handsets. In the CMS-based testing case, the HMM phonemes are trained by the CMS-processed clean speech data. By subtracting each cepstral mean value from each testing data(adaptation 1), the phrase accuracy is 74.7% for the ordinary analog telephone handsets, and 42.0% for the cordless telephone handsets. To compare with the result of HMM-SH, we attempt to subtract the cepstral mean of the same adaptation data from the testing data(adaptation 2). The phrase accuracy is 72.6% for the ordinary analog telephone handsets, and 38.6% for the cordless telephone handsets. These results show that the result of CMS is better than that of HMM-SH(without decomposition of noise HMM) in noisy telephone channel.

Table 4 shows the comparison with the matched condition for one ordinary analog telephone handset. In the case of the HMM-TELE(matched handset) which are trained by the speech through only one kind of the ordinary analog telephone handset, the performance is 86.6% for the same ordinary analog telephone handset. However, the phrase accuracy with the HMM-SHN is 80.1% for the same analog telephone handset with one adaptation sentence. There is the difference of the phrase accuracy between the HMM decomposition and the matched condition.

5. CONCLUSION

We have evaluated the performance of the model adaptation by the previously proposed HMM decomposition method[1] on the telephone speech recognition. The average phrase recognition accuracy with the clean speech HMMs is 60.9% for the ordinary analog telephone handsets, and 19.6% for the cordless telephone handsets. The average phrase recognition accuracy with the CMS HMMs is 74.7% for the ordinary analog telephone handsets, and 42.0% for the cordless telephone handsets. By the HMM decomposition method, the average phrase recognition accuracy is improved to 78.1% for the ordinary analog telephone handsets, and 50.5% for the cordless telephone handsets. These results show the HMM decomposition method is able to improve the performance. However, in the matched condition, the average phrase recognition accuracy is 77.7% for the ordinary analog telephone handsets, and 61.0% for the cordless telephone handsets. Therefore, the further improvement of the HMM adaptation method would be necessary for the cordless telephone speech.

6. REFERENCES

1. T.Takiguchi, S.Nakamura, Q.Huo, K.Shikano, "Adaptation of Model Parameters by HMM Decomposition in Noisy Reverberant Environments", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp.155-158, Apr.1997.
2. S.F.Boll,"Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. on ASSP*, Vol. ASSP-27, No.2, 1979.
3. B.Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *J. Acoust. Soc. Amer.*, Vol. 55, pp.1304-1312, 1974.
4. A.Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition", Ph.D Dissertation, ECE Department, CMU, Sept. 1990.
5. M.G.Rahim and B.-H.Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 1, pp.19-30, 1996.
6. M.Morishima, T.Isobe, N.Koizumi, "Phonetically Balanced Cepstrum Mean Normalization", *Acoust. Soc. America and Acout. Soc. Japan Third Joint Meeting*, pp.1105-1108, Dec. 1996.
7. S.Nakamura, T.Takiguchi, K.Shikano, "Noise and room acoustics distorted speech recognition by HMM composition", *Proc. ICASSP-96*, 1996, pp.69-72.
8. A.P.Varga, R.K.Moore, "Hidden Markov model decomposition of speech and noise", *Proc. ICASSP-90*, 1990, pp.845-848.
9. M.J.F.Gales, S.J.Young, "An improved approach to the hidden Markov model decomposition of speech and noise", *Proc. ICASSP-92*, 1992, pp.233-236.
10. M.J.F.Gales, S.J.Young, "PMC for speech recognition in additive and convolutional noise", CUED-F-INFENG-TR154, 1993.
11. F.Martin, K.Shikano, Y.Minami, "Recognition of noisy speech by composition of hidden Markov models", *Proc. EUROSPEECH-93*, 1993, pp.1031-1034.
12. A. Sankar and C.-H. Lee, "Robust speech recognition based on stochastic matching", *Proc. ICASSP-95*, 1995, pp.121-124.
13. V.Abrash, A.Sankar, H.Franco, M.Cohen, "Acoustic adaptation using transformations of HMM parameters", *Proc. ICASSP-96*, 1996, pp.729-732.
14. Y.Minami, S.Furui, "A maximum likelihood procedure for a universal adaptation method based on HMM composition", *Proc. ICASSP-95*, 1995, pp.129-132.
15. Y.Minami, S.Furui, "Adaptation method based on HMM composition and EM algorithm", *Proc. ICASSP-96*, 1996, pp.327-330.