FAST ADAPTATION OF ACOUSTIC MODELS TO ENVIRONMENTAL NOISE USING JACOBIAN ADAPTATION ALGORITHM

Yoshikazu YAMAGUCHI *Satoshi TAKAHASHI* *Shigeki SAGAYAMA*

NTT Human Interface Laboratories
1-1 Hikari-no-Oka, Yokosuka-shi, Kanagawa, 239 JAPAN
Tel: +81–468–59–2944 Fax: +81–468–55–1054
E-mail: {yamaguch,taka,saga}@nttspch.hil.ntt.co.jp

10.21437/Eurospeech.1997-543

ABSTRACT

This paper describes Jacobian adaptation (JA) of acoustic models to environmental noise and its experimental evaluation. JA is based on a "noise adaptation" idea, which is acoustic model adaptation from initial noise A to target noise B, and uses Jacobian matrices to relate changes in environmental noise with changes in the "speech+noise" acoustic model. It is experimentally shown that JA performs well compared with existing techniques such as HMM composition, particularly when only a short sample (shorter than 1 sec) of the target noise is given, and that JA is very advantageous in terms of computational cost. Moreover, this paper describes JA used in combination with noise spectral subtraction and shows that improving SNR by spectral subtraction leads to higher efficiency.

1. INTRODUCTION

In real applications of speech recognition, mismatch between training and testing environments often occurs, because environmental conditions may vary from time to time (e.g., mobile applications) or with a place (e.g., telephone applications). This results in a serious degradation of performance.

The best answer to this problem may be to retrain acoustic models with corrupted speech data observed in the test environment. However, the retraining requires many computations and a considerable amount of noise data, so it is not reasonable for real applications. HMM composition techniques, such as PMC [1] and NOVO [2], reduce the mismatch by combining a speech model and a noise model trained with data observed in the testing environment. These techniques require fewer computations and less noise data than the retraining, but they are insufficient for real-time acoustic model adaptation.

On the other hand, like speaker adaptation from initial speaker A to target speaker B, it is reasonable to consider the possibility of acoustic model adaptation from initial noise A to target noise B. Based on this "noise adaptation" idea, we proposed a fast model adaptation technique based on Jacobian matrices [3]. This technique is low in computation cost and requires only a small amount of noise data for acoustic model adaptation in comparison with NOVO (equivalently PMC).

In this paper, we focus on the performance of this method in adapting cepstrum parameters only and in adapting both cepstrum and delta cepstrum parameters. These were tested while changing noise spectral shapes, noise levels (SNR), and observation lengths of the target noise. Next, we tested various noise changes and investigated the range of noise changes that this method can handle. Finally, we evaluated the combination of this method and spectral subtraction (SS). We can expect further improvements by SS, because the performance of acoustic model adaptation to environmental noise, such as HMM composition and also our proposed method, depends strongly on the SNR.

2. JACOBIAN ADAPTATION (JA)

2.1. JA of Cepstra

Recently, the authors proposed a fast algorithm for acoustic model adaptation to environmental noise [3]. Since cepstrum parameters in "speech+noise" models are non-linear functions of the cepstrum of the background noise, a small change, $\Delta \boldsymbol{C}_{S+N}$, in "speech+noise" cepstrum, $\boldsymbol{C}_{S+N}$, is related to a small change, $\Delta \boldsymbol{C}_N$, in the noise cepstrum, $\boldsymbol{C}_N$, by a Jacobian matrix such that

$$\Delta \boldsymbol{C}_{S+N} = \frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N} \Delta \boldsymbol{C}_N. \tag{1}$$

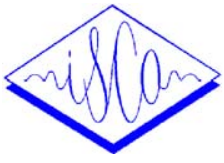
The Jacobian matrix is obtained as follows (See [3] for the detailed derivation).

$$\frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N} = \boldsymbol{F}^* \frac{\boldsymbol{N}}{\boldsymbol{S}+\boldsymbol{N}} \boldsymbol{F} \tag{2}$$

where $\boldsymbol{S}$, $\boldsymbol{N}$, and $\boldsymbol{S}+\boldsymbol{N}$ represent the clean speech spectrum, the initial noise spectrum, and the initial "speech+noise" spectrum, respectively. $\frac{\boldsymbol{N}}{\boldsymbol{S}+\boldsymbol{N}}$ is the element-wise division of the vector $\boldsymbol{N}$ by $\boldsymbol{S}+\boldsymbol{N}$. $\boldsymbol{F}$ and $\boldsymbol{F}^*$ are the Fourier transform matrix and its transposed complex conjugate. The matrices can be calculated in the training phase.

2.2. JA of Delta Cepstra

Similar to adapting cepstrum mean vectors of HMMs, delta cepstrum mean vectors can also be compensated by Jacobian matrices. Here we consider delta



EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

cepstra of the time derivative of cepstra. Denoting the time derivative of $\boldsymbol{C}$ by $\dot{\boldsymbol{C}}$, the formulations for adapting delta cepstrum parameters are the following equations.

$$\Delta\dot{\boldsymbol{C}}_{S+N} = \frac{\partial\dot{\boldsymbol{C}}_{S+N}}{\partial\boldsymbol{C}_N}\Delta\dot{\boldsymbol{C}}_N \tag{3}$$

$$\frac{\partial\dot{\boldsymbol{C}}_{S+N}}{\partial\boldsymbol{C}_N} = -\boldsymbol{F}^*\frac{\boldsymbol{N}\dot{\boldsymbol{S}}}{(\boldsymbol{S}+\boldsymbol{N})^2}\boldsymbol{F} \tag{4}$$

where $\dot{\boldsymbol{S}}$ represents the time derivative of the clean speech spectrum. $\boldsymbol{N}\dot{\boldsymbol{S}}$ is the element-wise multiplication of the vector $\boldsymbol{N}$ by $\dot{\boldsymbol{S}}$. In these formulas, we ignore $\frac{\partial\dot{\boldsymbol{C}}_{S+N}}{\partial\dot{\boldsymbol{C}}_N}$ and $\dot{\boldsymbol{N}}$, because we assume that the mean of the delta cepstrum of the noise signal is 0.

From the above point-to-point relationship in cepstrum and delta cepstrum domains, the noise adaptation algorithm is derived for the mean vectors and covariance matrices of individual distributions [3].

2.3. The JA Algorithm for Continuous Mixture HMMs

The noise adaptation procedure by JA for the mean vectors of continuous mixture HMMs is summarized as follows. (The covariance matrices are left unchanged in this paper.)

[**Training Phase**]

Step 1: Assume initial noise A and train initial "speech+noise" models. Alternatively, the initial models can be composed by PMC or NOVO from existing clean speech HMMs and the assumed noise.

Step 2: Calculate a Jacobian matrix for each of the mean vectors in the initial "speech+noise" HMMs.

[**Recognition Phase**]

Step 3: Observe target noise B for adaptation (e.g., just before the target speech) and obtain the noise cepstral means.

Step 4: Update all cepstrum and delta cepstrum mean vectors in the initial "speech+noise" HMMs from Jacobian matrices and from the differences of parameters between the assumed and the observed noises.

All we have to do for adaptation in the recognition phase is simply to multiply a $p \times p$ matrix by a pth-order vector ($p = 17$ in the following experiments) and add to the mean vector for each of the Gaussian distributions. Thus the "speech+noise" model can be adapted instantaneously after the target noise observation.

3. COMBINATION OF JA AND SPECTRAL SUBTRACTION (SS)

To enhance JA, we combine JA and noise spectral subtraction (SS). (Hereinafter, this combination is called "SS-JA".) SS formulation is as follows [4].

$$\hat{\boldsymbol{S}} = (\boldsymbol{S}+\boldsymbol{N}) - \alpha\,\bar{\boldsymbol{N}}$$
$$\tilde{\boldsymbol{S}} = \begin{cases} \hat{\boldsymbol{S}} & \text{if } \hat{\boldsymbol{S}} > \beta\,(\boldsymbol{S}+\boldsymbol{N}) \\ \beta\,(\boldsymbol{S}+\boldsymbol{N}) & \text{otherwise} \end{cases} \tag{5}$$

where $\bar{\boldsymbol{N}}$ is the average noise spectrum, and enhanced speech $\tilde{\boldsymbol{S}}$ is the input for training and recognition. α is an overestimation factor and β is a flooring factor.

In JA, it is necessary to observe the changes of noise component directly. In the SS procedure, however, the noise component is given by the overestimation and underestimation components of the noise in the above procedure, so these estimation errors are not directly observed. Therefore, we roughly obtained the noise component including the estimation errors using the following procedure in SS-JA.

$$\hat{\boldsymbol{N}} = \boldsymbol{N} - \alpha_N\,\bar{\boldsymbol{N}}$$
$$\tilde{\boldsymbol{N}} = \begin{cases} \hat{\boldsymbol{N}} & \text{if } \hat{\boldsymbol{N}} > \beta_N\,\boldsymbol{N} \\ \beta_N\,\boldsymbol{N} & \text{otherwise} \end{cases} \tag{6}$$

where $\tilde{\boldsymbol{N}}$ is the overestimation and underestimation error components of the noise spectrum. α_N and β_N are the overestimation and flooring factors. In this paper, we adopted NOVO models combined with SS (SS-NOVO) as the initial models used in SS-JA. SS-NOVO also uses the noise estimation procedure in Eq. (6) when the noise model is trained.

Table 1. CPU time for adaptation (not including acoustic analysis).

phase	JA		NOVO
	cep	cep+dcep	
training	2,216 ms	8,033 ms	4,416 ms
recognition	149 ms	349 ms	5,066 ms

(measured on Sun SPARCstation20)

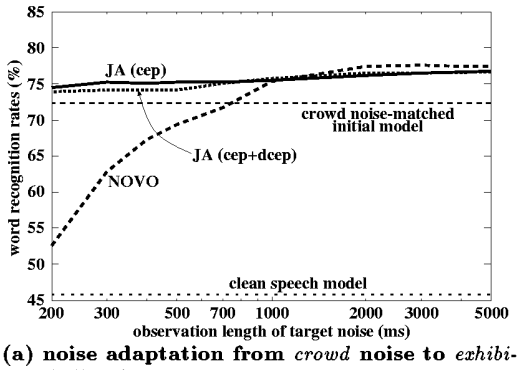
4. EXPERIMENTAL EVALUATION

4.1. Conditions

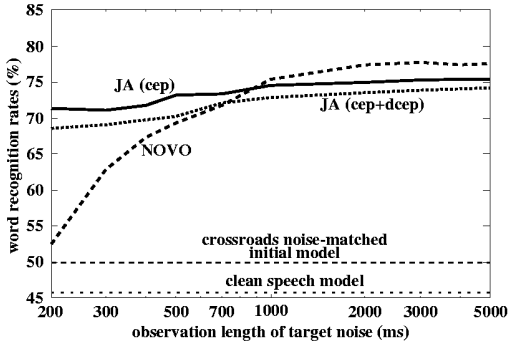
JA was experimentally evaluated on isolated 400-word speech recognition and compared with NOVO (equivalently PMC). The test speech data were noise-corrupted data of 100 city names uttered by 13 speakers. In word recognition, we used 33-order feature vectors: 16-order LPC cepstrum vectors, 16-order delta cepstrum vectors and a delta log power. In adaptation, we added a log power. Initial "speech+noise" models used in JA were composed by NOVO. Also, initial noise model was trained with 60-sec noise data in all experiments. The JA algorithm was applied to cepstrum parameters only and to both cepstrum and delta cepstrum parameters, and the performance was compared. In figures and tables, these are indicated as JA (cep) and JA (cep+dcep).

4.2. Fundamental Performance of JA

Table 1 shows the CPU time for JA and NOVO. In terms of adaptation after target noise is observed,



(a) noise adaptation from *crowd* noise to *exhibition hall* noise.



(b) noise adaptation from *crossroads* noise to *exhibition hall* noise.

Figure 1. Word recognition rates for various observation lengths of the target noise at 10 dB SNR.

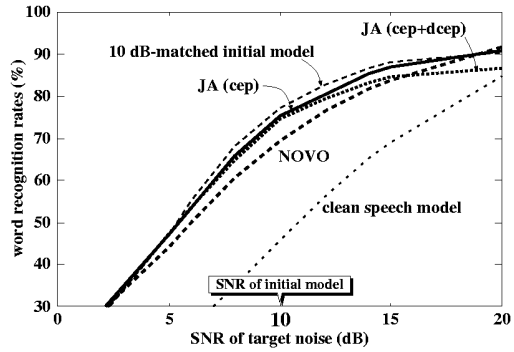


Figure 2. Word recognition rates for changes of SNR from initial 20 dB; observation length of target noise is 500 ms.

JA required approximately 1/34 the computational cost of NOVO, and 1/15 when additionally adapting the delta cepstrum parameters. These results imply that JA is suitable for instantaneous (online real-time) acoustic model adaptation to environmental noise.

Figure 1 shows word recognition rates of noise adaptation from a different noise to *exhibition hall* noise for various observation lengths of target noise. The results show JA performed better than NOVO for short lengths of the target noise, even for 200 ms. The decrease of the performance of NOVO with short lengths of target data is caused by the estimation error of noise variance.

Table 2. Word recognition rates of JA from 7 typical initial noise models, the target noise being 500 ms at 10 dB SNR; ERR denotes the error reduction rate.

initial noise	target noise	initial model (%)	JA (%)	ERR (%)
computer room		11.2	61.1	56.2
factory		18.8	71.8	65.3
passing trains	*exhi-*	42.5	70.7	49.0
crossroads	*bition*	50.0	73.2	46.4
in car	*hall*	55.4	68.7	29.8
railway station		66.2	73.3	21.0
crowd		72.4	75.2	10.1

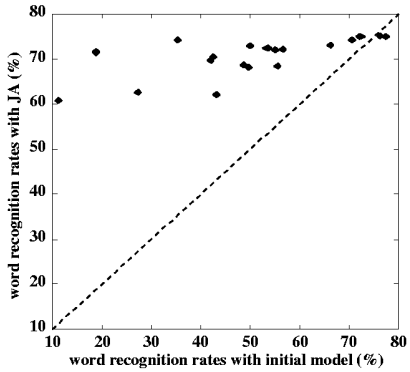


Figure 3. Relationship between word recognition rates with noise-mismatched initial model and with JA from 19 initial noise models to the target of 500-ms *exhibition hall* noise at 10-dB SNR.

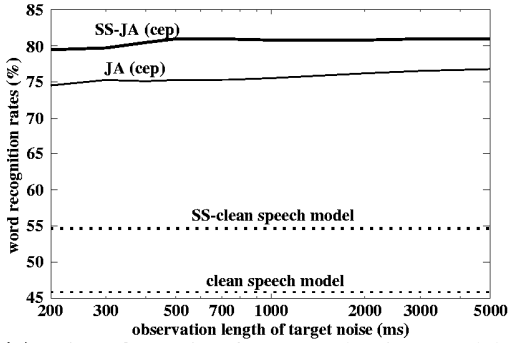
In Fig. 1(a), adaptation of both cepstrum and delta cepstrum parameters improved the recognition rate when the noise data length was relatively long (i.e., 800 – 3,000 ms). In Fig. 1(b), however, adaptation of both parameters showed even lower performance over the whole noise length than adaptation of cepstrum parameters only. We consider that adaptation of delta cepstrum parameters improves the recognition rate when the initial noise and target noise are close and the long target noise is obtained.

Figure 2 shows word recognition rates for changes of SNR. The initial model was trained with 10-dB *exhibition hall* noise. The target noises used in JA were 500-ms *exhibition hall* noise with the SNR ranging from 0 dB to 20 dB. We couldn't find a significant advantage compared with using the initial model in mismatched SNR.

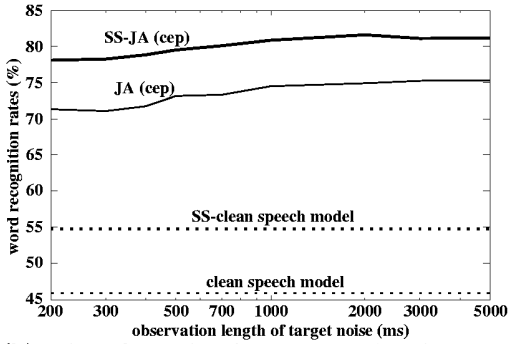
4.3. Performance for Various Noise Changes

JA is based on the idea of "noise adaptation" where initial noise A is adapted to target noise B. To investigate how large differences are allowed in JA from the initial noise A to the target noise B, we tested 19 various initial noises (*crowd*, *crossroads*, *passing trains*, etc.), the target noise being fixed as *exhibition hall*.

Table 2 and Fig. 3 show word recognition rates with



(a) noise adaptation from *crowd* noise to *exhibition hall* noise.



(b) noise adaptation from *crossroads* noise to *exhibition hall* noise.

Figure 4. Word recognition rates combining with SS for various observation lengths of the target noise at 10 dB SNR.

the noise-mismatched initial model and that with JA (in adaptation cepstrum only). Seven of the noise changes are listed in Table 2 as examples and the relationships between them in all 19 noise changes are plotted in Fig. 3. In these results, the higher word recognition rate with the initial model implies the initial noise is closer to the target noise.

Since JA is based on first order Taylor approximation, it might be supposed that the performance of JA is limited within or near the linearity range and JA is able to handle only small changes in noise. These results indicate that, as the target noise becomes further from the initial noise, recognition performance with JA is degraded gradually. However, the recognition performance was significantly improved by JA, even if the recognition rates with the initial models were low; for example, the recognition rate improved from 11.2 % to 61.1 % for noise change from *computer room* noise to *exhibition hall* noise. We can see that JA works even for large changes in noise.

4.4. Combination of JA with SS

Figure 4 and 5 show the respective improvements obtained by combining JA with SS (denoted "SS-" in the figures) under the same conditions in Fig. 1 and 2. The average noise spectrum used in SS was estimated from 20-frame data just before the utterance. Initial "speech+noise" models used in SS-JA were composed by SS-NOVO. These result confirm that the combination of JA and SS provided a large improvement in recognition rate.

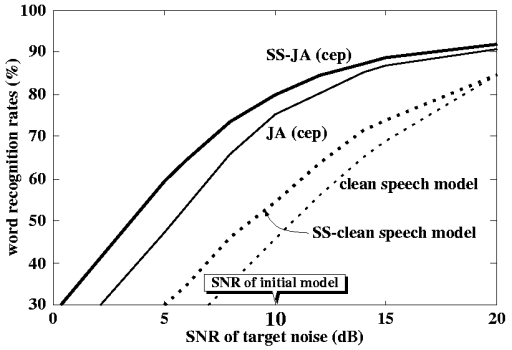


Figure 5. Word recognition rates combining with SS for changes of SNR; observation length of target noise is 500 ms.

5. CONCLUSION

We have presented an experimental evaluation of fast adaptation of acoustic models to environmental noise based on JA. The advantages of this method are that it requires only a small amount of noise data and reduces computation cost compared with NOVO (equivalently PMC). Although adaptation of delta cepstrum parameters improved the recognition performance in the limited cases, the above advantages enable acoustic models to be adapted to fluctuating environmental noise in each speaker's utterance when only a short noise is observed (e.g., between the guidance sentences) before the utterance. Moreover, it was found that performance could be improved by combining of JA and SS.

The evaluation of JA for various noise adaptations demonstrated it could handle large changes in noise, even if the initial noise was not close to the target noise. As the target noise becomes further from the initial noise, however, recognition performance with JA is degraded gradually. Our future work will include multiple initial "speech+noise" models from which the closest noise condition to the target noise is selected before JA is applied.

REFERENCES

[1] M. J. F. Gales and S. J. Young, "An Improved Approach to the Hidden Markov Model Decomposition of Speech And Noise," *Proc. ICASSP92*, pp. 233–236, 1992.

[2] F. Martin, K. Shikano and Y. Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models," *Proc. Eurospeech93*, pp. 1031–1034, 1993.

[3] S. Sagayama, Y. Yamaguchi, S. Takahashi and J. Takahashi, "Jacobian Approach to Fast Acoustic Model Adaptation," *Proc. ICASSP97*, pp. 835–838, 1997.

[4] J. A. N. Flores and S. J. Young, "Adapting a HMM-Based Recognizer for Noisy Speech Enhanced by Spectral Subtraction," *Proc. Eurospeech93*, pp. 829–832, 1993.