Jacobian Approach to Joint Adaptation to Noise, Channel and Vocal Tract Length

Shigeki Sagayama[1,2], *Yutaka Kato*[2], *Mitsuru Nakai*[2] *and Hiroshi Shimodaira*[2]

[1] *The University of Tokyo*, Bunkyo-ku, Tokyo 113-8656 Japan
[2] *Japan Advanced Institute of Science and Technology*, Tatsu-no-kuchi, Ishikawa 923-1292 Japan
E-mail: sagayama@hil.t.u-tokyo.ac.jp, {sagayama,ykatou,mit,sim}@jaist.ac.jp

Abstract

This paper describes the Jacobian approach to simultaneously adapting acoustic models to unknown noise, channel and vocal tract length from a supervised adaptation data. As has been both theoretically and experimentally shown, Jacobian adaptation is one of most efficient methods for model adaptation if the target condition is close to the initial condition. It utilizes the linear relationship in the neighbor of the initial condition which in turn can be used in decomposition of multiple factors. The analytic relationship between noise, channel, vocal tract length and the observed cepstrum is linearized using the Jacobian matrices. Least squares fit gives the estimates of noise, channel and vocal tract stretch parameters. Experimental evaluation gave a significant improvement to the recognition accuracy.

1 Introduction

Jacobian approach[4, 5] (first conceived in 1992, experimentally evaluated in 1995, and presented in Japanese in 1996) is a method of adapting initial acoustic models under an initial condition, A, to a target condition, B, assuming that the difference between the two conditions is relatively small. This framework is motivated by speaker adaptation from the initial speaker, A (or speaker-independent), to the target speaker, B. Based on this idea, we introduced an approach that uses a Jacobian matrix. Being a straight forward linear computation, Jacobian adaptation is computationally very efficient and is one of most suitable solutions in practical applications of speech recognition where environmental conditions vary from time to time (e.g., mobile applications) or with each usage (e.g., telephone applications) and cause acoustic model mismatch resulting in a serious degradation of performance. Retraining acoustic models using the actual noise environment is not feasible in most cases in terms of training data amount and computational time constraint. Composing mathematically a noisy speech model from acoustic models of clean speech and noise such as Parallel Model Composition (PMC)[1] and Noise-Voice Composition (NOVO)[2] is often computationally too expensive to follow in real-time the instantaneous changes in noise spectrum and level.

In recent further research results using Jacobian adaptation such as initial model selection[6] and transformation of Jacobian matrices[7, 8], this approach has been made not only more computationally efficient but also more effective in its performance to a given noisy condition.

Differential approach including Jacobian adaptation has a high potential in model adaptation. It can be applied to any factors whose relationship to the feature vectors is analytically known. If the initial and target conditions are close enough for good linear approximation, this approach is applicable to an arbitrary function, not being limited within a specific function such as in the Vector Taylor Series[3] which is related to Jacobian adaptation in a sense that first-order approximation of Taylor series is Jacobian. The Jacobian approach opens up a new non-parametric analytic paradigm of acoustic model adaptation other than existing probabilistic (e.g., Bayesian approaches) or parametric approaches (e.g., MLLR). In addition, by introducing any parametric model, Jacobian approach turns to be a parametric adaptation method.

In this paper, we discuss an extension of the Jacobian approach to simultaneous adaptation to unknown noise and channel, and also to adaptation to vocal tract length.

If there is a small change in linear channel characteristics, this approach is also easily applicable. Given a small change included in the adaptation problem, it is also solved in the same way as in the noise case. It is further extended to adaptation to vocal tract length changes since there is an analytic relationship between vocal tract length stretch and the resulted spectrum.

It is often the case that channel and noise change simultaneously and can not be observed separately. This arouses a joint estimation problem of channel and noise. This may be solved with a computationally expensive iterative algorithm. If, however, the changes of channel and noise characteristics between the initial and target conditions are small, Jacobian approach can make the problem linear so that simple least squares fit gives a computationally efficient algorithm.

In this paper, we extend the formulation to include the channel problem. We also discuss the future directions of this approach.

2 Jacobian Adaptation to Noise and Channel

2.1 Formulating Jacobian Adaptation

If a vector variable, $\boldsymbol{C}_Y$, is an analytic function of other variables, $\boldsymbol{C}_S$, $\boldsymbol{C}_N$ and $\boldsymbol{C}_H$, namely,

$$\boldsymbol{C}_Y = \boldsymbol{\Psi}(\boldsymbol{C}_S, \boldsymbol{C}_N, \boldsymbol{C}_H), \tag{1}$$

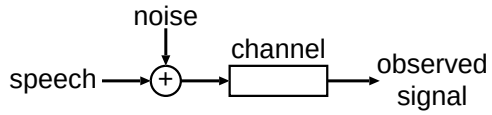


Figure 1: The model of speech, noise and channel.

a small change in $\boldsymbol{C}_Y$ caused by small changes in $\boldsymbol{C}_S, \boldsymbol{C}_N$ and $\boldsymbol{C}_H$ is given by

$$\Delta \boldsymbol{C}_Y = \frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{C}_S}\Delta \boldsymbol{C}_S + \frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{C}_N}\Delta \boldsymbol{C}_N + \frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{C}_H}\Delta \boldsymbol{C}_H \quad (2)$$

regardless of the meaning of variables $\boldsymbol{C}_S, \boldsymbol{C}_N$ and $\boldsymbol{C}_H$. We call $\frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{C}_S}$, $\frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{C}_N}$ and $\frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{C}_H}$ Jacobian matrices whose (i, j) component is the derivative of the ith component of $\boldsymbol{\Psi}$ in respect to the jth component of $\boldsymbol{C}$, i.e., $\frac{\partial \Psi_i}{\partial C_j}$.

The above mathematical relationship holds if n-dimensional vectors $\boldsymbol{S}$, $\boldsymbol{N}$, $\boldsymbol{H}$, and $\boldsymbol{Y}$ represents clean speech spectrum, additive noise spectrum, multiplicative channel characteristics (transfer function in the power spectral domain), and the resulted composite speech spectrum, respectively, and $\boldsymbol{C}_S, \boldsymbol{C}_N, \boldsymbol{C}_H$ and $\boldsymbol{C}_Y$ represent their corresponding cepstrum vectors.

When noise and channel conditions represented by $\boldsymbol{C}_N$ and $\boldsymbol{C}_H$ ("Condition A") change into $\boldsymbol{C}_N + \Delta \boldsymbol{C}_N$ and $\boldsymbol{C}_H + \Delta \boldsymbol{C}_H$ ("Condition B") with the speech spectrum fixed, the composite cepstrum also changes into:

$$\boldsymbol{C}_Y + \Delta \boldsymbol{C}_Y = \boldsymbol{\Psi}(\boldsymbol{C}_S,\ \boldsymbol{C}_N + \Delta \boldsymbol{C}_N,\ \boldsymbol{C}_H + \Delta \boldsymbol{C}_H) \quad (3)$$

where $\Delta \boldsymbol{C}_H$ is given by Eq.(2) and $\Delta \boldsymbol{C}_S = 0$. This is the basic idea of Jacobian adaptation to a new condition.

2.2 Jacobian for Noise and Channel

The relationship among n-dimensional vectors, $\boldsymbol{S}$, $\boldsymbol{N}$, $\boldsymbol{H}$, and $\boldsymbol{Y}$ shown in Fig. 1 is given by

$$\boldsymbol{Y} = \boldsymbol{H}(\boldsymbol{S} + \boldsymbol{N}) \quad (4)$$

in the linear spectral domain. These relations also hold in a particular case where $\boldsymbol{C}_S, \boldsymbol{C}_N, \boldsymbol{C}_H$ and $\boldsymbol{C}_Y$ represent cepstra of speech, noise, channel and resulted composite signal, respectively, as shown in In the linear spectral domain, holds Equivalently, in the cepstral domain, we have[1]

$$\boldsymbol{C}_Y = \boldsymbol{F}^* \Big[\log \big\{ \exp(\boldsymbol{F}\boldsymbol{C}_S) + \exp(\boldsymbol{F}\boldsymbol{C}_N) \big\} \Big] + \boldsymbol{C}_H \quad (5)$$

where $\boldsymbol{F}$ is the Fourier transform matrix[2] and $\boldsymbol{F}^*$ is the transposed complex conjugate of $\boldsymbol{F}$ that $\boldsymbol{F}^*\boldsymbol{F} = \mathbf{1}$.

The signal-to-noise ratio is not treated separately; the noise power is included in $\boldsymbol{N}$. Since these cepstra are related to the respective spectra by

$$\begin{array}{ll} \log \boldsymbol{Y} = \boldsymbol{F}\boldsymbol{C}_Y, & \log \boldsymbol{S} = \boldsymbol{F}\boldsymbol{C}_S, \\ \log \boldsymbol{N} = \boldsymbol{F}\boldsymbol{C}_N, & \log \boldsymbol{H} = \boldsymbol{F}\boldsymbol{C}_H \end{array} \quad (6)$$

[1]For vectors $\mathbf{a}$ and $\mathbf{b}$, we define $\mathbf{ab} = (a_1 b_1, \cdots, a_n b_n)^T$, $\mathbf{a}/\mathbf{b} = (a_1/b_1, \cdots, a_n/b_n)^T$, $\exp \mathbf{a} = (\exp a_1, \cdots, \exp a_n)^T$, and $\log \mathbf{a} = (\log a_1, \cdots, \log a_n)^T$. One can regard vectors as diagonal matrices for consistency with matrix arithmetic.

[2]Fourier transform matrix, $\boldsymbol{F}$, is not necessarily a square matrix. It depends on the resolution of frequency and the number of cepstral points (quefrencies).

It should be noted that the Fourier transform is essentially the cosine transform in real symmetric spectrum cases, namely, $F_{ij} = \cos \frac{2ij\pi}{n}$ and the computation amount can be reduced to one quarter by handling the positive frequency only. One implementation is

$$F_{ik} = \cos \frac{i(k + 0.5)\pi}{N} \quad (7)$$

where N stands for the number of frequency points. Cepstrum can be simply replaced by MFCC in the above formulation where the corresponding spectrum is replaced by mel-frequency warped spectrum.

If these changes are small, the resulted change, $\Delta \boldsymbol{C}_Y$, is

$$\Delta \boldsymbol{C}_Y = \frac{\partial \boldsymbol{C}_Y}{\partial \boldsymbol{C}_N}\Delta \boldsymbol{C}_N + \Delta \boldsymbol{C}_H \quad (8)$$

according to Eq. (2).

The Jacobian matrix is easily calculated at the initial condition A:

$$\begin{aligned} \boldsymbol{J}_C \ &\equiv\ \frac{\partial \boldsymbol{C}_Y}{\partial \boldsymbol{C}_N} = \frac{\partial \boldsymbol{C}_Y}{\partial \log \boldsymbol{Y}} \frac{\partial \log \boldsymbol{Y}}{\partial \boldsymbol{Y}} \frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{N}} \frac{\partial \boldsymbol{N}}{\partial \log \boldsymbol{N}} \frac{\partial \log \boldsymbol{N}}{\partial \boldsymbol{C}_N} \\ &=\ \boldsymbol{F}^* \frac{\mathbf{1}}{\boldsymbol{H}(\boldsymbol{S} + \boldsymbol{N})} \boldsymbol{H}\boldsymbol{N}\boldsymbol{F} = \boldsymbol{F}^* \frac{\boldsymbol{N}}{\boldsymbol{S} + \boldsymbol{N}} \boldsymbol{F} \end{aligned} \quad (9)$$

Thus, if differences between the initial and observed conditions, A and B, is found in the cepstrum domain, i.e., $\Delta \boldsymbol{C}_N$ and $\Delta \boldsymbol{C}_H$, the composite cepstrum, $\boldsymbol{C}_Y + \Delta \boldsymbol{C}_Y$, is approximately computed by Eq. (8).

Eq.(9) implies that any analytic relationship can be decomposed into a product and/or sum of simple Jacobian matrices. Just like Bayesian networks, we can consider "Jacobian newtworks" which include multiple causes and intermediate results connected in a network structure.

2.3 Least Mean Squares Estimation

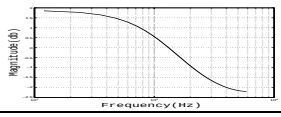
Noise and channel differences, however, can not be observed separately in most cases such as in telephone speech recognition. It is therefore necessary to simultaneously estimate the noise and channel spectral characteristics from the observed speech signal.

Suppose that we observe noisy speech containing multiple phonemes through a channel. Note that Eq. (4) commonly holds at all different $\boldsymbol{S}$ corresponding to multiple phonemes. If M hidden states of the model are aligned (e.g., by Viterbi algorithm) to the input speech, we have a set of M different simultaneous relations derived from Eq. (4). Assuming that the target condition B is relatively close to the initial condition A, we have a set of M linear equations with additional errors derived from Eq. (8) as follows:

$$\left\{ \begin{array}{lcl} \Delta \boldsymbol{C}_Y^{(1)} &=& \boldsymbol{J}_C^{(1)} \Delta \boldsymbol{C}_N + \Delta \boldsymbol{C}_H + \boldsymbol{\epsilon}^{(1)} \\ \Delta \boldsymbol{C}_Y^{(2)} &=& \boldsymbol{J}_C^{(2)} \Delta \boldsymbol{C}_N + \Delta \boldsymbol{C}_H + \boldsymbol{\epsilon}^{(2)} \\ & \vdots & \\ \Delta \boldsymbol{C}_Y^{(M)} &=& \boldsymbol{J}_C^{(M)} \Delta \boldsymbol{C}_N + \Delta \boldsymbol{C}_H + \boldsymbol{\epsilon}^{(M)} \end{array} \right. \quad (10)$$

Thus, if a small amount of input speech with unknown noise and channel is given with its phonetical transcription, we can obtain the joint estimate of $\Delta \boldsymbol{C}_N$ and $\Delta \boldsymbol{C}_H$ by simple

Table 1: Experimental conditions for noise and channel simultaneous adaptaion.

Speech DB	ATR Speech DB A-set (5240 words)
Speakers	MAU (male), FFS (female)
Training	2650 words (odd-numbered)
Testing	655 words (from even-numbered)
Features	13 MFCCs + 13 ΔMFCCs
Models	3-state, 4-mixture, CD phone HMMs
Noise A	Exhibition Hall (10dB SNR)
Noise B	Crowd (0, 10, 20, 30 dB)
Channel A	Flat
Channel B	Simulated (shown right) 

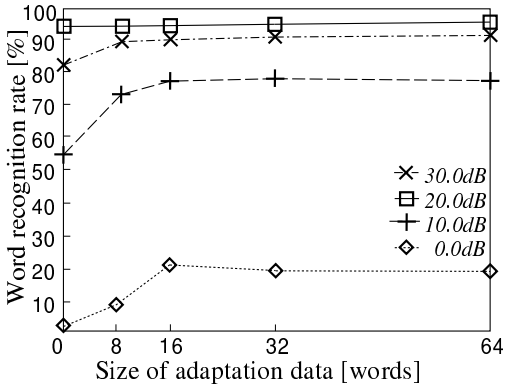


Figure 2: A typical example of simultaneous noise and channel adaptation of 10-dB SNR initial models to given noise and channel conditions.

least squares estimation minimizing the sum of squares of error terms, $\sum |\boldsymbol{\epsilon}^{(i)}|^2$. This procedure can be iterated for more accurate Viterbi alignment using the adapted models.

This process estimates only two vectors, $\Delta \boldsymbol{C}_N$ and $\Delta \boldsymbol{C}_H$, and is advantageous from the statistical estimation point of view. Once the linear decomposition into $\Delta \boldsymbol{C}_N$ and $\Delta \boldsymbol{C}_H$ are estimated, Eq. (8) is applied to all mean vectors of the model for adapting the whole model.

2.4 Experimental Evaluation

For experimental evaluation of simultaneous adaptation to noise and channel, speaker-dependent isolated-word recognition was performed with artificially added environmental noise and simulated channel under conditions shown in Table 1. A typical result is shown in Fig. 2. Table 2 shows the averaged results from adaptation experiments across different noise sources and SNRs ranging from 0dB to 30dB. A significant error reduction around 10% is observed for only 8 words given for supervised adaptation, while more data do not give more improvement. This is due to the quite limited number of free parameters to estimate in the least squares fit in the linear decomposition of mixed factors.

Table 2: Simple average recognition and error reduction rates in simultaneous adaptation across mismatched noise conditions.

Adapt. Data Size (wds)	0	8	16	32	64
Ave. Recog. Rate (%)	33.6	40.6	40.9	40.3	40.0
Error Reduc. Rate (%)	——	10.6	10.6	10.1	9.8

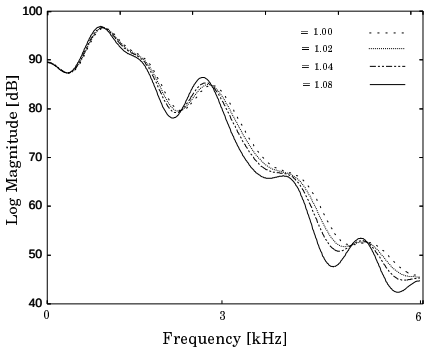


Figure 3: Resulted spectrum of the cepstral-domain frequency stretching by Jacobian approximation

3 Jacobian Adaptation to Vocal Tract Length

3.1 Vocal Tract λ-stretched Cepstrum

The Jacobian approach can be extended to include some aspects of speaker differences. If the vocal tract length becomes λ-times longer, the corresponding speech spectrum changes from $f(\omega)$ into $f(\lambda\omega)$. We can again calculate the Jacobian matrix $\boldsymbol{J}_\lambda$ of resulted cepstrum in respect to λ. (This matrix has only one column and appears like a vector.)

Combining noise, channel, and vocal tract length factors, we can express the small changes in the composite cepstrum as follows using small changes in noise cepstrum, channel cepstrum, and the vocal length stretch coefficient:

$$\Delta \boldsymbol{C}_Y = \boldsymbol{J}_N \Delta \boldsymbol{C}_N + \Delta \boldsymbol{C}_H + \boldsymbol{J}_\lambda \Delta\lambda \tag{11}$$

In Eqs.(6), if we replace the Fourier Transform $\boldsymbol{F}$ with a λ-stretched Fourier transform $\boldsymbol{F}^\lambda$, or, if we use

$$F^\lambda_{ik} = \cos \frac{\lambda i(k+0.5)\pi}{N} \tag{12}$$

instead of Eq.(7), the λ-stretched speech spectrum $\tilde{\boldsymbol{S}}$ is given by

$$\log \tilde{\boldsymbol{S}} = \boldsymbol{F}^\lambda \boldsymbol{C}_S. \tag{13}$$

The λ-stretched cepstrum $\tilde{\boldsymbol{C}}_S$ is thus expressed as

$$\tilde{\boldsymbol{C}}_S = \boldsymbol{F}^{-1} \boldsymbol{F}^\lambda \boldsymbol{C}_S \tag{14}$$

whose i-th component is given by

$$\tilde{C_{Si}} = \sum_{j=1}^{N} F^{-1}_{ij} \sum_{k=1}^{p} F^\lambda_{jk} C_{Sk} \tag{15}$$

from which the i-th component of its Jacobian in respect to λ is derived by differentiating it by λ as

$$J_{\lambda i} = \sum_{j=1}^{N} F^{-1}_{ij} \sum_{k=1}^{p} \frac{-j(k+0.5)\pi}{N} G_{jk} C_{Sk} \tag{16}$$

where matrix G represents the sine transform.

Fig. 3 shows the spectral domain display of cepstral-domain frequency stretching with Jacobian linear approximation by $\boldsymbol{J}_\lambda \Delta\lambda$. From this figure, we can understand that the approximation works only near $\lambda = 1.0$ but has a limitation as λ leaves 1.0.

3.2 Joint Adaptation to Three Factors

As preliminary evaluation of frequency stretching, we calculated appropriate $lambda$ values by comparing spectral peak frequencies of initial model (condition A) and target (condition B) spectra derived from observed cepstra. Among several ways of finding and applying λ values to the adaptation procedure in a speaker adaptation experiments across 4 speakers, averaged λ values extracted from vowels and applied to all phonemes yielded 4% recognition error reduction with 2 training words and 5% with 4, 8, and 16 words.

In joint adaptation to noise, channel and speaker, the following procedure was taken:

- Preparation (before adaptation):
  1. Train the initial noise model ($\boldsymbol{C}_N^A$)
  2. Train the noisy speech model ($\boldsymbol{C}_Y^A$)
  3. Calculate Jacobian matrices ($\boldsymbol{J}_N, \boldsymbol{J}_\lambda$)
- Adaptation (supervised):
  4. Train the target noisy speech model ($\boldsymbol{C}_Y^B$) using a small amount of supervised training data through Viterbi time alignment.
  5. Observe the difference ($\Delta\hat{\boldsymbol{C}}_Y^B$) between conditions A and B.
  6. LMS-estimate the differences of noise and channel ($\Delta\boldsymbol{C}_N$, $\Delta\boldsymbol{C}_N$) between A and B.
  7. Find $\Delta\lambda$ by comparing $\boldsymbol{C}_Y^A$ and $\boldsymbol{C}_Y^B$.
  8. Calculate $\Delta\tilde{\boldsymbol{C}}_Y{}^A$ using $\Delta\lambda$ and $\boldsymbol{J}_\lambda$.
  9. Obtain the adapted model ($\tilde{\boldsymbol{C}}_Y{}^B$) combining the changes $\Delta\boldsymbol{C}_N, \Delta\boldsymbol{C}_H$ and $\Delta\boldsymbol{C}_\lambda$.
- Recognition:
  10. Recognize the speech data in condition B using the adapted model.

A typical results of applying this joint adaptation technique to different noise, channel and speaker is shown in Fig. 4 where 16 LPC cepstrum coefficients and their time derivatives were used as features. It was experimentally found that joint adaptation gives more effective results than the sum of separate adaptation to separate factors. A possible explanation is that speaker differences include not only vocal tract length stretch but also multiplicative (and possibly additive) spectrum differences, and that the mixed problem of noise, channel and speaker adaptation is better solved by joint adaptation in respect to all of three factors.

4 Conclusion

This paper introduced linear decomposition of noise and channel differences in mismatched condisions using a Jacobian formulation. Joint adaptation to noise, channel and vocal tract length (frequency stretching) mismatches and demonstrated with experimental results was also discussed. Future works include overall least mean squares joint estimation of noise, channel, and vocal tract length formulated in a similar way as described for the noise and channel case.

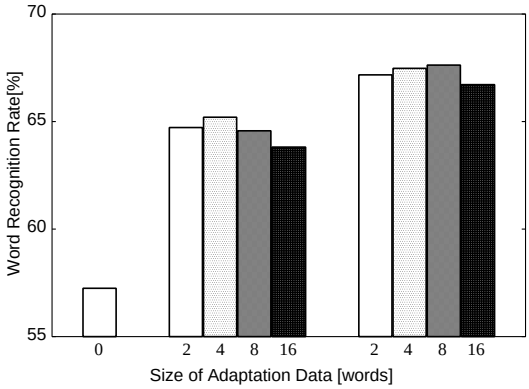


Figure 4: A typical results of word speech recognition of no adaptation (left), joint adaptation to additive (noise), multiplicative (channel) factors (center) and adaptation to additive, multiplicative and frequency stretching (vocal tract length) factors. Initial condition (*A*): models were trained with 3 speakers (mht, mms, mmy), flat channel, 30dB SNR noise "in factory"; Testing condition (*B*): speaker "mau", low-pass channel, 10dB SNR noise "crossroads".

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, et al., "Recognition of Noisy Speech by Using the Composition of Hidden Markov Models," *Proc. 1992 Fall ASJ Conf.*, 1-7-10, 1992.

[3] P. J. Moreno, B. Raj, and R. Stern, "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," *Proc. ICASSP96*, pp. 733–736, 1996.

[4] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi: "Jacobian Approach to Fast Acoustic Model Adaptation," *Proc. ICASSP97*, Vol. 2, pp. 835–838, 1997.

[5] Y. Yamaguchi, S. Takahashi, and S. Sagayama, "Fast Adaptation of Acoustic Models to Environmental Noise Using Jacobian Adaptation Algorithm," *Proc. Eurospeech 97*, pp. 2051–2054, 1997.

[6] H. Shimodaira, T. Akae, M. Nakai and S. Sagayama, "Jacobian Adaptation of HMM with Initial Model Selection for Noisy Speech Recognition," *Proc. ICSLP2000 (Beijing)*, pp.1003-1006, 2000.

[7] C. Cerisara, L. Rigazio, R. Boman and J.-C. Junqua, "Transformation of Jacobian Matrices for Noisy Speech Recognition," *Proc. ICSLP2000 (Beijing)*, pp. 369-372, 2000.

[8] R. Sarikaya and J. Hansen, "Improved Jacobian Adaptation for Fast Acoustic Model Adaptation in Noisy Speech Recognition," *Proc. ICSLP2000 (Beijing)*, pp. 702–705, 2000.