JACOBIAN JOINT ADAPTATION TO NOISE, CHANNEL AND VOCAL TRACT LENGTH

Hiroshi Shimodaira, Nobuyoshi Sakai, Mitsuru Nakai, and Shigeki Sagayama

Japan Advanced Institute of Science and Technology
School of Information Science
Tatsunokuchi, Ishikwaka 923-1292 Japan

The University of Tokyo
Graduate School of Information Science and Technology
Hongo, Bunkyo-ku, Tokyo 113-8654 Japan

ABSTRACT

A new Jacobian approach that linearly decomposes the composite of additive noise, multiplicative noise (channel transfer function) and speaker's vocal tract length, and adapts the acoustic model parameters simultaneously to these factors is proposed in this paper. Due to the fact that these factors non-linearly degrade the observed features for speech recognition, existing approaches fail to adapt the acoustic models adequately. Approximating the nonlinear operation by a linear model enables to employ the least square error estimation of the factors and adapt the acoustic model parameters with small amount of speech samples. Speech recognition experiments on ATR isolated word database demonstrate significant reduction of error rates, which supports the effectiveness of the proposed scheme.

1. INTRODUCTION

Acoustic features for speech recognition are degraded by number of factors such as background noise, transfer-functions of communication channels and so on. Some of these factors may affects the observed features non-linearly. For example, additive noise in the power spectrum domain degrades the cepstral features non-linearly.

There have been number of researches conducted to adapt the acoustic model parameters to the target environments in the real world where speech recognition is carried out. In spite of these efforts, none of the adaptation techniques outperforms the acoustic models that have been trained sufficiently in the target environment. This fact suggests us another scenario of adaptation in which (1) we assume we had a set of acoustic models trained in the same environment with the target environment, (2) the characteristics of the target environment will slightly change in accordance with the time goes by or as the speaker moves in the mobile environment, (3) only small amount of modification of the model parameters is enough to follow the environmental change. In such situations where real-time parameter updating is necessary, direct use of modern well-known techniques, PMC [1] and MLLR [2] for example, is not adequate due to the computation complexity. Jacobian adaptation (JA) [3, 4, 5, 6] is one of the solutions for this problem because everything is carried out in the cepstral domain and hence no transformation of feature vectors into spectral domain is needed.

The basic concept of the Jacobian adaptation is to model the observed features as a analytic function that may be nonlinear, and approximate the function into a linear form so that the acoustic model parameters can be adapted in the feature domain of the acoustic models without transforming the parameters into other domain such as linear spectral domain.

2. JACOBIAN ADAPTATION

2.1. Formulation of Jacobian Adaptation

In the Jacobian approach, observed vector variable is given by an analytic function of some variables [1]. In the present study, we assume the observed cepstral vector C_Y is a function of vector variables, C_S, C_N and C_H, and a scalar variable λ, namely,

$$C_Y = \Psi(C_S, \lambda, C_N, C_H) \tag{1}$$

If C_S, C_N, C_H and λ change by a very small quantity as ΔC_S, ΔC_N, ΔC_H and $\Delta\lambda$, corresponding small change of C_Y is expressed by

$$\Delta C_Y = \frac{\partial \Psi}{\partial C_S}\Delta C_S + \frac{\partial \Psi}{\partial \lambda}\Delta\lambda + \frac{\partial \Psi}{\partial C_N}\Delta C_N + \frac{\partial \Psi}{\partial C_H}\Delta C_H \tag{2}$$

We call $\partial\Psi/\partial C_S, \partial\Psi/\partial C_N$, and $\partial\Psi/\partial C_H$ Jacobian matrices whose (i,j) component is the derivative of the ith component of Ψ with respect to the jth component of C, i.e., $\partial\Psi_i/\partial C_j$, and we call $\partial\Psi/\partial\lambda$ a Jacobian vector with respect to λ.

Since the above mathematical relationship holds regardless the meaning of variables C_S, C_N, C_H, C_Y and λ, one can assume that the first four vectors are the cepstral vectors corresponding to the spectral vectors S_S, S_N, S_H and S_Y which represent clean speech, additive noise and multiplicative channel characteristics (transfer function in the

[1] The Vector Taylor Series [7] employs a similar approach.

0-7803-7402-9/02/$17.00 ©2002 IEEE

power spectral domain) and the resultant composite speech spectrum that we observe. The last variable λ reflects a factor of speaker's vocal tract length, which will be discussed in detail in the following section. Fig. 1 depicts the observing system assumed in this paper.

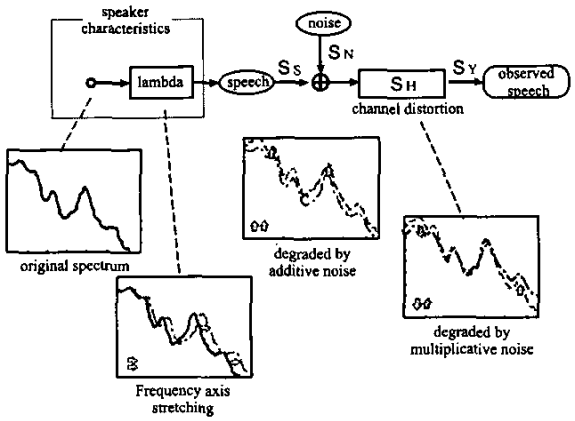


Fig. 1. Observing system: the model of speech, vocal tract length, noise and channel

In case that the noise and channel conditions represented by C_N and C_H ("Condition A") change into $C_N+\Delta C_N$ and $C_H+\Delta C_H$ ("Condition B") with the clean speech spectrum fixed, the composite cepstrum also changes into:

$$C_Y+\Delta C_Y = \Psi(C_S,\ \lambda,\ C_N+\Delta C_N,\ C_H+\Delta C_H) \quad (3)$$

where ΔC_Y is given by Eq.(2) with $\Delta C_S = 0$, $\Delta\lambda = 0$.

Since we have got the above adaptation formula, we can adapt the acoustic model parameters, i.e., the mean vectors of output distributions of HMMs, after observing a small amount of audio samples for adaptation in a new condition B. Note that the adaptation is performed with very small computation complexity because (1) every operation is done in the same feature domain with the acoustic model parameters, i.e., the cepstral domain in this case, and hence no transformation of the features to the spectral domain is needed, (2) Jacobian matrices are calculated once in the initial condition A regardless the target condition B.

This is the basic idea of Jacobian adaptation to a new condition.

2.2. Vocal Tract λ-stretched Cepstrum and its Jacobian

In our previous work of Jacobian adaptation [4], the noise and channel factors have been considered. The speaker's vocal tract length is newly employed as the third factor (variable) in the present study.

We assume that if the vocal tract becomes λ times longer in length then corresponding power spectrum changes from $S(\omega)$ to $S(\lambda\omega)$. We call this $S(\lambda\omega)$ a lambda-stretched spectrum, and call the corresponding cepstrum a λ-stretched cepstrum.

The relationship between the cepstrum and the power spectrum is expressed as

$$\log \boldsymbol{S} = \boldsymbol{F}\boldsymbol{C} \quad (4)$$

where $\boldsymbol{F}$ is the Fourier transform matrix. Since the power spectrum is real and symmetric, the Fourier transform can be simplified with the discrete cosine transform (DCT) matrix whose (i,k) element is given by

$$F_{ik} = \cos\frac{i(k+0.5)\pi}{N}. \quad (5)$$

With a little calculation, the above relationship can be applied to the case of λ-stretched spectrum and its relationship to the original cepstrum is written as

$$\log \tilde{\boldsymbol{S}} = \boldsymbol{F}^{\lambda}\boldsymbol{C}_S. \quad (6)$$

where $\boldsymbol{F}^{\lambda}$ denotes another DCT matrix including the λ stretching of frequency axis, whose (i,k) element is given by

$$F^{\lambda}_{ik} = \cos\frac{\lambda i(k+0.5)\pi}{N}. \quad (7)$$

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

$$\tilde{\boldsymbol{C}}_S = \boldsymbol{F}^{-1}\boldsymbol{F}^{\lambda}\boldsymbol{C}_S \quad (8)$$

whose i-th component is given by

$$\tilde{C}_{Si} = \sum_{j=1}^{N} F^{-1}_{ij}\sum_{k=1}^{p} F^{\lambda}_{jk}C_{Sk}. \quad (9)$$

Since $\tilde{\boldsymbol{C}}_S$ is now represented as an analytic function of λ, the i-th component of the Jacobian with respect to λ is derived as

$$(\boldsymbol{J}_\lambda)_i \approx \sum_{j=1}^{N} F^{-1}_{ij}\sum_{k=1}^{p}\frac{-j(k+0.5)\pi}{N}G_{jk}C_{Sk} \quad (10)$$

where the matrix G represents the sine transform, and we further assumed $\lambda \approx 1$. As a result, we obtain the following expression if we discard other factors, noise and channel, that affect the observed $\boldsymbol{C}_Y$:

$$\Delta\boldsymbol{C}_Y = \frac{\partial\Psi}{\partial\lambda}\Delta\lambda = \boldsymbol{J}_\lambda\Delta\lambda. \quad (11)$$

2.3. Jacobians for Noise and Channel

The relationship among the n-dimensional vectors, $\boldsymbol{S}_S$, $\boldsymbol{S}_N$, $\boldsymbol{S}_H$, and $\boldsymbol{S}_Y$ in the linear spectral domain shown in Fig. 1 is given by

$$\boldsymbol{S}_Y = \boldsymbol{S}_H(\boldsymbol{S}_S + \boldsymbol{S}_N). \quad (12)$$

On the other hand, in the cepstral domain, the relationship for the corresponding vectors is rewritten as [2]

$$C_Y = F^* \Big[\log \big\{ \exp(FC_S) + \exp(FC_N) \big\} \Big] + C_H \quad (13)$$

where F is the Fourier transform matrix and F^* is the transposed complex conjugate of F that $F^*F = 1$.

If the changes of C_N and C_H are small, the resulted change, ΔC_Y, is denoted by

$$\Delta C_Y = \frac{\partial C_Y}{\partial C_N} \Delta C_N + \Delta C_H \quad (14)$$

according to Eq. (2).

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

$$\begin{aligned} J_N &\equiv \frac{\partial C_Y}{\partial C_N} \\ &= \frac{\partial C_Y}{\partial \log S_Y} \frac{\partial \log S_Y}{\partial S_Y} \frac{\partial S_Y}{\partial S_N} \frac{\partial S_N}{\partial \log S_N} \frac{\partial \log S_N}{\partial C_N} \\ &= F^* \frac{1}{S_H(S_S+S_N)} S_H S_N F = F^* \frac{S_N}{S_S+S_N} F \end{aligned} \quad (15)$$

Thus, if the differences between the initial and observed conditions, A and B, are found in the cepstrum domain, i.e., ΔC_N and ΔC_H, the composite cepstrum, $C_Y + \Delta C_Y$, is approximately computed by Eq. (14).

3. JACOBIAN JOINT ADAPTATION

In the previous section, we have shown that the small change of the observed cepstrum C_Y caused by the small changes of C_λ, C_N and C_H is approximated as a linear combination written as:

$$\Delta C_Y = J_\lambda \Delta\lambda + J_N \Delta C_N + \Delta C_H. \quad (16)$$

Thus the adaptation of acoustic models can be carried out if we can observe these changes, $\Delta\lambda$, $J_N\Delta C_N$, ΔC_H. In the real world environment, however, noise, channel and vocal tract length will change simultaneously, and thus one can not easily tell the exact changes of the factors. For example, if one could assume that noise alone changed among the three factors, then he/she could observe ΔC_N directly from the pause or silent signal regions where no speech sounds exists. However, in case that noise and channel change simultaneously, he/she could not recover ΔC_N and ΔC_H from the observed signal directly. To tackle this problem we employ the least square estimation approach to decompose the factors from the speech signals.

To put the problem into the least square estimation problem that is solvable, we need to obtain a set of equations of (16) which are hopefully independent each other. To that end, in the target environment B, we are going to observe a sequence of speech signal which contains number of different phonemes. Note that for each observed sound corresponding to each phoneme Eq. (16) commonly holds. Therefore, in case that there are M hidden states in all in the set of acoustic models, we will have the following system of linear equations with the observation error term $\epsilon^{(i)}$:

$$\begin{cases} \Delta C_Y^{(1)} &= J_\lambda^{(1)} \Delta\hat{\lambda} + J_N{}^{(1)} \Delta\widehat{C_N} + \Delta\widehat{C_H} + \epsilon^{(1)} \\ \Delta C_Y^{(2)} &= J_\lambda^{(2)} \Delta\hat{\lambda} + J_N{}^{(2)} \Delta\widehat{C_N} + \Delta\widehat{C_H} + \epsilon^{(2)} \\ &\vdots \\ \Delta C_Y^{(M)} &= J_\lambda^{(M)} \Delta\hat{\lambda} + J_N{}^{(M)} \Delta\widehat{C_N} + \Delta\widehat{C_H} + \epsilon^{(M)} \end{cases} \quad (17)$$

where $\Delta C_Y^{(i)}$ stands for the change of C_Y, when the initial condition A was changed to the target condition B, which was observed in the speech region where the i-th hidden state was assigned. In the implementation issue, we assume that we know the text uttered, i.e., the phoneme sequence, and time alignment between the observed speech signal and the hidden states of phoneme models is properly performed by time alignment algorithm such as Viterbi algorithm.

Solving the problem to minimize the sum of error terms, $\Sigma \left|\epsilon^{(i)}\right|^2$, yields the estimates, $\Delta\widehat{C_N}, \Delta\widehat{C_H}, \Delta\hat{\lambda}$. Once we have got these estimates, Eq. (16) is applied to adapt all the mean vectors of the HMMs to the target condition.

It should be noted that giving a linear combination formulation like Eq. (16) with respect to the factors by the Jacobian approximation is the essential idea to achieve the simultaneous adaptation of the acoustic models to the factors.

4. EXPERIMENTS

4.1. Experimental Conditions

The proposed Jacobian joint adaptation scheme was evaluated using a speech database of isolated words, in which experimental condition is shown in Table 1.

4.2. Experimental Results

Fig. 2 and Fig. 3 show word recognition rates for different target SNR conditions, when the number of adaptation words was 16. Comparing with the results when only the two factors, noise and channel, are adapted, the proposed joint adaptation demonstrates better recognition performance, though the improvement is not larger than the case of noise and channel adaptation. This is because, as other studies have shown frequency axis stretching is not as effective as other speaker adaptation schemes. Moreover, it will be more likely in the proposed scheme that the model mismatch caused by different speakers has been adapted by not only the factor of vocal tract length but also the other two factors, noise and channel.

[2] Here we assume that the logarithmic and exponential functions also apply to a vector, in which corresponding operation is applied to each element of the vector.

Table 1. Experimental conditions for noise and channel simultaneous adaptation.

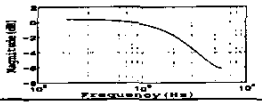
Speech DB	ATR Speech DB A-set (5240 words)
Training	2650 words (odd-numbered)
Testing	655 words (from even-numbered)
Features	16 LPC-CEPs + 16 Δ LPC-CEPs
Models	3-state, 3-mixture, CD phone HMMs
Speaker A	MAU (male) or FFS (female)
Speaker B	MHT (male) or FMS (female)
Noise A	Station yard or Car inside (10dB SNR)
Noise B	Intersection or Factory (0, 10, 20, 30 dB)
Channel A	Flat
Channel B	Simulated (shown right) 

Fig. 4 indicates that 8 or 16 words are enough to achieve the maximum error reduction performance of the proposed scheme. If one employs more sophisticated decomposition algorithm to estimate the changes of the factors in Eq. (16), one may be able to reduce the size of data for adaptation.

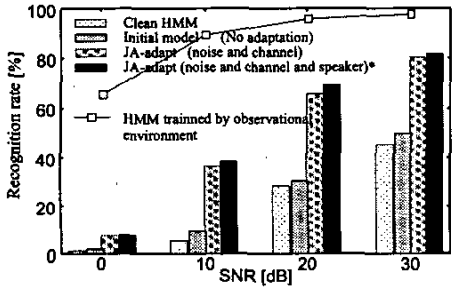


Fig. 2. Recognition rates for each SNR (male speaker)

5. CONCLUSIONS

This paper introduced a linear decomposition scheme of noise, channel and vocal tract length differences in the mismatched concisions, and simultaneous adaptation of the acoustic model parameters using a Jacobian formulation. Experiments on isolated word recognition task, though very preliminary, showed that the proposed joint adaptation scheme significantly reduced the word error rates with small number of speech samples for adaptation.

The proposed joint adaptation scheme is quite general and it is not limited to the case that the model mismatch is caused by noise, channel and vocal tract length, but it is also applicable to the case when the mismatch occurs due to any reasons containing any additive, multiplicative and frequency stretching disturbances.

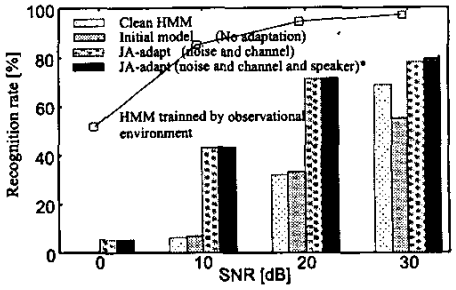


Fig. 3. Recognition rates for each SNR (male speaker)

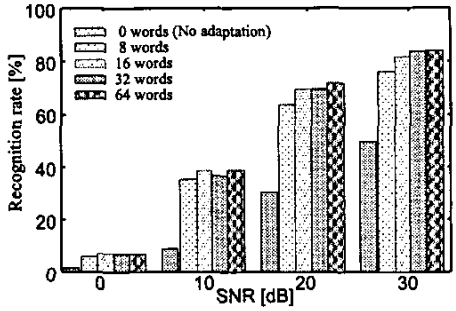


Fig. 4. Recognition rates vs. adaptation data size

6. REFERENCES

[1] M. J. F. Gales and S. J. Young, "An improved approach to the hidden markov model decomposition of speech and noise," in *Proc. ICASSP*, 1992, vol. 1, pp. 233–236.

[2] C. J. Legatter 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.

[3] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *Proc. ICASSP*, 1997, pp. 835–838.

[4] Hiroshi Shimodaira, Toshihiko Akae, Mitsuru Nakai, and Shigeki Sagayama, "Jacobian Adaptation of HMM with Initial Model Selection for Noisy Speech Recognition," in *Proc. ICSLP*, 2000, pp. 1003–1006.

[5] Alex Acero, Li Deng, Trausti Kristjansson, and Jerry Zhang, "HMM adaptation using vector taylor series for noisy speech recognition," in *Proc. ICSLP*.

[6] Ruhi Sarikaya and John H.L. Hansen, "Improved jacobian adaptation for fast acoustic model adaptation in noisy sapeech recognition," in *Proc. ICSLP*, 2000, p. [01550].

[7] Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern, "A Vector Taylor Series Approach for Environment-independent Speech Recognition," in *Proc. ICASSP*, 1996, vol. 2, pp. 733–736.