JACOBIAN ADAPTATION OF HMM WITH INITIAL MODEL SELECTION FOR NOISY SPEECH RECOGNITION

Hiroshi SHIMODAIRA Yutaka KATO Toshihiko AKAE Mitsuru NAKAI Shigeki SAGAYAMA

School of Information Science
Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa, 923–1292 JAPAN
http://iipl.jaist.ac.jp/index.html

ABSTRACT

An extension of Jacobian Adaptation (JA) of HMMs for degraded speech recognition is presented in which appropriate set of initial models is selected from a number of initial-model sets designed for different noise environments. Based on the first order Taylor series approximation in the acoustic feature domain, JA adapts the acoustic model parameters trained in the initial noise environment A to the new environment B much faster than PMC that creates the acoustic models for the target environment from scratch. Despite the advantage of JA to PMC, JA has a theoretical limitation that the change of acoustic parameters from the environment A to B should be small in order that the linear approximation holds. To extend the coverage of JA, the ideas of multiple sets of initial models and their automatic selection scheme are discussed. Speaker-dependent isolated-word recognition experiments are carried out to evaluate the proposed method.

1. INTRODUCTION

Acoustic models show poor recognition performance when they were trained in a different environment from the recognition environment. Since such mismatch often occurs in degraded speech recognition, adapting the model parameters to the target environment is indispensable to achieve high recognition performance. Furthermore, in case of telephone speech recognition including the one in mobile environment, both noise and channel characteristics change every moment so that fast online adaptation is necessary.

Although HMM composition method such as Parallel Model Combination (PMC) [1] and NOVO [2] successfully makes HMMs for any degraded speech from both clean speech and noise HMMs, the algorithm is not suitable for the online adaptation due to its inefficiency in computational complexity and large amount of adaptation data. On the other hand, Jacobian Adaptation (JA) [3, 4] requires adaptation data of very short period of time (0.5 seconds for example) and it needs much less computational cost than that of PMC. This advantage of JA to PMC comes from two reasons. One is that JA adapts the model parameters in the same domain with the model parameters such as cepstrum coefficients by using a linear approximation based on the Taylor series expansion, while PMC does most of its computation in the linear-scale power spectrum domain. The second reason, though it has not been confirmed yet, comes from the fact that JA adapts the model parameters that have been trained sufficiently in the initial environment (env-A) to the target recognition environment (env-B) hopefully similar to env-A. Therefore small amount of adaptation data is enough for the adaptation. On the other hand, PMC composes the acoustic models for env-B from both the clean-speech models and the noise model of env-B. Since the clean-speech models are usually far from those for env-B, PMC needs large amount of adaptation data of env-B to compose HMMs for env-B.

Since JA employs a linear approximation based on the 1st order Taylor series of a non-linear function around the points of the model parameters for env-A, the approximation accuracy deteriorates as the target env-B becomes far from env-A. Such ill situation can be avoided by setting up multiple sets of initial models trained in the various noise environments and choosing the most appropriate set based on a similarity criterion.

2. JACOBIAN ADAPTATION

We at first assume that the clean speech is degraded by a additive noise and corrupted through a transfer channel as is shown in Fig. 1[1]. In the power spectrum domain, assuming that the speech and noise are statistically independent each other, the power spectrum of the observed degraded speech $S_Y(\omega)$ is expressed by

$$S_Y(\omega) = S_H(\omega)(S_S(\omega) + S_N(\omega)) \quad (1)$$

where $S_H(\omega), S_S(\omega), S_N(\omega)$ denote the power spectrum of the channel transfer function, the clean speech, and the additive noise, respectively. Discretizing the radial frequency, let S_Y, S_H, S_N be the vectors in R^n corresponding to $S_Y(\omega), S_H(\omega), S_N(\omega)$.

The relationship between a vector S in the power spectrum domain and its corresponding vector C in the cepstrum domain is expressed by

$$S = \exp(FC) \quad (2)$$

[1]One can assume another different observation modeling where the clean speech is transfered through a channel at first and the noise is added afterward. JA can be applied even in such a situation with small modification to the mathematical formulations described in this paper.

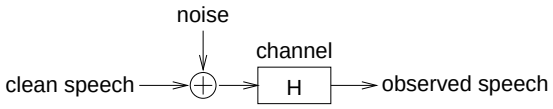

Figure. 1: observation model

where F denotes the Fourier transform matrix, and exp represents the exponential operation on each vector element. The expression of (1) is now rewritten in the cepstrum domain by

$$C_Y = F^{-1}[\log\{\exp(FC_S) + \exp(FC_N)\}] + C_H. \quad (3)$$

One can see from the expression that C_Y (the observed speech in the cepstrum domain) is given as a non-linear function of C_S, C_N and C_H.

In case that the observation condition changes sufficiently small and we can assume small changes of both C_S, C_N and C_H, the new observed speech $\tilde{C}_Y^{new}$ is approximated by the first order Taylor series as

$$\begin{aligned} \tilde{C}_Y^{new} &= C_Y + \Delta C_Y & (4) \\ &= C_Y + \frac{\partial C_Y}{\partial C_S}\Delta C_S + \frac{\partial C_Y}{\partial C_N}\Delta C_N + \Delta C_H. & (5) \end{aligned}$$

As a result, the non-linearity disappears and C_Y is expressed as a linear function of C_S, $C_N$and C_H.

To simplify the problem, we assume that both C_H and C_S remain unchanged and only the C_N changes so that the relationship is denoted by

$$\tilde{C}_Y^{new} = C_Y + \frac{\partial C_Y}{\partial C_N}\Delta C_N \quad (6)$$

where $\partial C_Y / \partial C_N$ is the Jacobian matrix and denoted by J_N here.

The Jacobian matrix J_N is calculated by

$$\begin{aligned} J_N &\equiv \frac{\partial C_Y}{\partial C_N} \\ &= \frac{\partial C_Y}{\partial \log S_Y}\frac{\partial \log S_Y}{\partial S_Y}\frac{\partial S_Y}{\partial S_N}\frac{\partial S_N}{\partial \log S_N}\frac{\partial \log S_N}{\partial C_N}. \quad (7) \end{aligned}$$

The (i, j) element of J_N is given by

$$(J_N)_{ij} = \sum_k (F^{-1})_{ik}\frac{(S_N)_k}{(S_S)_k + (S_N)_k}(F)_{kj}. \quad (8)$$

It should be noted that calculating J_N does not require any information about the new environment but it needs just S_S and S_N, both of which are provided in the initial environment.

Though there is a preceding work [5] that also utilizes the Taylor series approximation, JA [6] was proposed independently with it, and its formulation and the approach to degraded speech recognition are different.

2.1. Jacobian Adaption of HMM

The former discussion on the Jacobian Adaptation deals with the point-to-point relationship of vectors in both the linear spectrum domain and cepstrum domain (Fig. 2). Special consideration should be given when one applies JA to the adaptation of

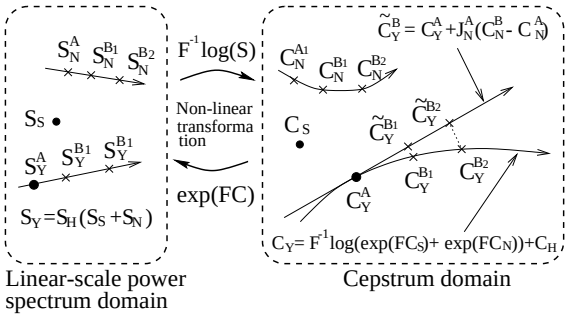

Figure. 2: Non-linear effect of the noise fluctuation on the cepstrum feature vectors of the degraded speech, and its linear approximation by the Taylor series expansion.

the HMM's stochastic parameters. For that, we assume that the variance of the distribution of C_Y is sufficiently small and stays within the effective rage of linear (Jacobian) approximation.

Although not only the mean vector of each distributions of HMMs but also the variance and those for delta-cepstrum can be adapted to the new environment, only the adaptation of the mean vector of C_Y is considered in this paper. This is because the adaptation of other parameters did not show significant improvement in the recognition performance in our previous report [3].

Denoting the mean vector of a distribution of HMMs by Mean[C_Y], each mean vector of distributions of HMMs trained in env-A can be adapted to env-B by

$$\text{Mean}[C_Y^B] = \text{Mean}[C_Y^A] + J_N^A \Delta\text{Mean}[C_N] \quad (9)$$

where the superscript denotes the environment.

In the framework of Jacobian Adaptation of HMMs, HMMs are firstly trained with the enough amount of data in the initial environment A (env-A), then adaptation to the recognition environment B (env-B) is done by observing a noise of very short period of time. We will describe below the three major processing phases of JA; training, adaptation, and recognition phase.

Training phase: (setting up the initial models)

Step 1 Train a noise HMM with the noise data in env-A.

Step 2 Train the acoustic HMMs with the degraded speech where noise is added to the clean speech in env-A. (Although PMC was used for this purpose in this research, the models can be trained directly with the degraded speech.)

Step 3 Calculate the Jacobian matrix J_N^A for each distribution of all the acoustic HMMs.

Adaptation phase:

Step 4 Observe small amount of noise signal in the testing environment (env-B) just before the utterance, and train the noise HMM to obtain Mean[C_N^B].

Step 5 Calculate the change of noise by

$$\Delta\text{Mean}[C_N] = \text{Mean}[C_N^B] - \text{Mean}[C_N^A]. \quad (10)$$

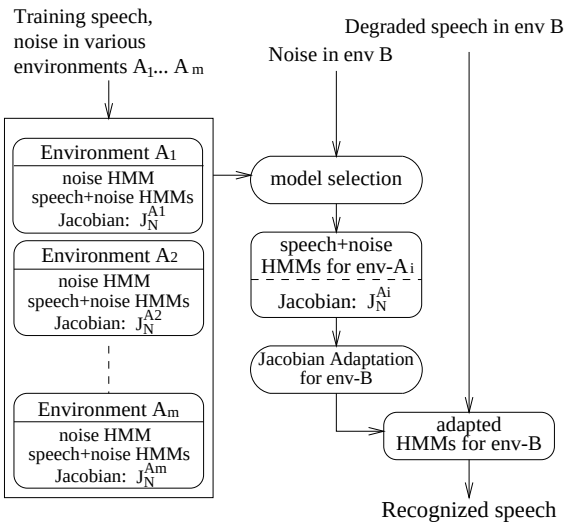

Figure. 3: JA based speech recognition using the multiple sets of initial models

Step 6 Adapts each mean vector of all the distributions using the expression (9).

Recognition phase:

Step 7 Using the adapted models, carry out speech recognition on the test speech data in env-B .

2.2. Multiple Sets of Initial Models and Model Selection

Since JA assumes that the background noise changes small enough to adapt the model parameters by means of linear approximation given by (6), the accuracy of the approximation deteriorates when the the linearity assumption does not hold due to the dissimilarity between the two environments. As a result, the coverage of adaptation of a single set of initial models against the possible noise fluctuation may be small. In order to make the coverage of the adaptation wider, we can set up multiple sets of initial models trained in the various noise conditions and choose the most fitting model for the recognition environment (Fig. 3).

As a distance measure for selecting the initial-model set among the multiple sets of initial models, we employed the Bhattacharyya distance between the noises of the environment A and B, of which formulation is given by,

$$D_{AB} = \frac{1}{8}(\mu^B - \mu^A)^t \left(\frac{\Sigma^B + \Sigma^A}{2}\right)^{-1} (\mu^B - \mu^A) + \frac{1}{2}\ln \frac{|(\Sigma^B + \Sigma^A)/2|}{|\Sigma^B|^{\frac{1}{2}}|\Sigma^A|^{\frac{1}{2}}}. \quad (11)$$

where μ and Σ represent the mean vector and the covariance matrix, respectively, and the superscript denotes the environment.

3. EXPERIMENTS

3.1. Experimental conditions

We conducted speaker-dependent isolated word speech recognition experiments in the following manner.

Initial model training speaker-dependent phone HMMs (3-states, 4-mixtures, context-independent HMM) for the initial noise environment A was composed using the PMC method [1]. For that, a set of clean-speech phone HMMs for each of 4 speakers (2-males, 2-females) was trained with the 2620 words speech data of ATR A-set database, and the noise HMM (1-state, single Gaussian) was trained with each of the 4 different noise sources (car, exhibition-hall, intersection, crowd) of 60 seconds of its duration. The clean-speech HMMs and the noise HMM were blended at 4 different SNRs, 0, 10, 20, 30dB, hence 16 different sets of degraded speech HMMs (4 noises * 4 SNRs) were prepared in all. Then the Jacobian matrix for each mean vector of distributions of HMM is calculated.

Testing data preparation The testing data for recognition experiment was made by computationally adding the noise signal (4 noise sources * 4 SNRs) to the clean speech signal of 655 words. The noise data comes from the same noise sources but different instances with the above.

Adaptation Before the recognition of each word in the target environment B, preceding noise of 0.5 seconds is observed to train the noise HMM and the difference between the mean vectors of the noise HMMs for both the env-A and env-B. Then each mean vector of all the Gaussian distributions of the initial HMMs is adapted to the env-B by the JA method.

Recognition Word recognition experiment was carried out for each of 4 speakers. In each test for a certain noise source and SNR, those initial-model sets of which noise source are the same with that of testing environment are excluded. As a result, 12 sets (3 noise sources * 4 SNRs = 12) of initial models were used for the initial-model selection.

3.2. Experimental results and discussion

Before discussing the results of the proposed method, Fig. 4 demonstrates the limitation of JA when only a single set of initial models was used. Each folded-line shows word recognition rates of the adapted HMMs that were trained at a certain SNR. We can see from the figure that a single set of adapted HMMs shows the highest recognition performance when the SNR of the initial environment coincides with the one of the testing environment in this case. But the performance decreases rapidly as the difference between the two SNRs becomes larger.

Fig. 5 shows recognition performance by the multiple sets of initial models (denoted by JA(12) in the figure), where the testing environment is the same with Fig. 4. Besides the performance by the proposed method, results by PMC and the average recognition rates by the single initial-model JA (denoted by JA(1)) with different training conditions (3 noise sources * 4 SNRs) are shown in the figure as well. Each vertical bar denotes the highest and lowest recognition rates among those by JA(1). Comparing with the

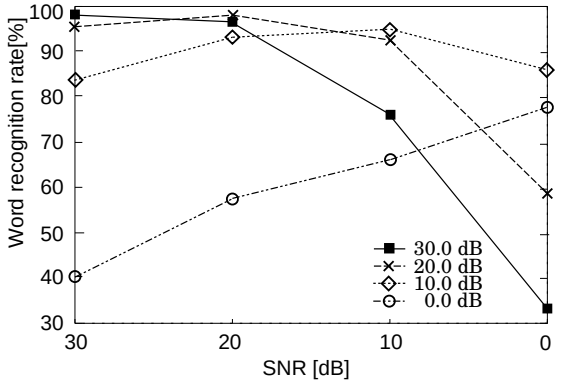

Figure. 4: Limitation of JA using a single set of initial models. (The initial environment: exhibition hall, the testing environment: car)

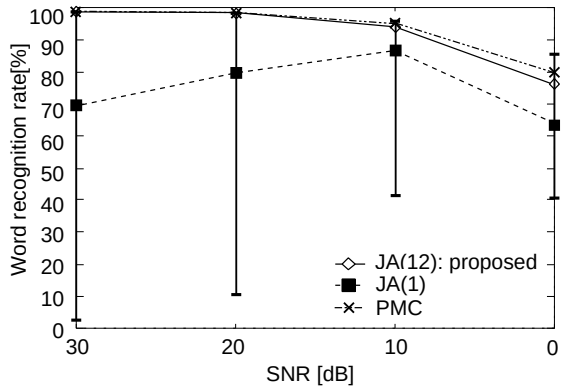

Figure. 5: Word recognition performance by the multiple sets of initial models. (testing environment: car)

former result by a single initial-model set, the proposed method dramatically improved the recognition performance that is almost equal to that of PMC.

The error reduction rates from the clean speech HMMs for the four different methods, JA(1), selection(12), proposed and PMC, are shown in Fig. 6, where the reduction rates denote the average rates of all the possible 12 conditions, “selection(12)” represents the case that only the initial model selection among the 12 initial sets was employed without adaptation. We can see that the proposed method achieved the comparable recognition performance with PMC when PMC used the 60 seconds noise data for adaptation while the proposed methods used only 0.5 seconds noise data.

4. CONCLUSION

In this paper, one of the theoretical limitations of Jacobian Adaptation of acoustic HMMs for degraded speech recognition has been relaxed by implementing the multiple sets of initial models

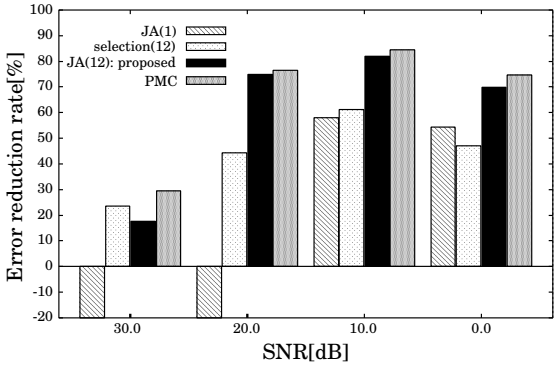

Figure. 6: Error reduction rates for ”JA(1)”: single initial-model set, ”selection(12)”: model selection without adaptation, ”JA(12)”: the proposed JA with multiple sets of initial models. The number in the parentheses denotes the number of initial-model sets

for various noise environments. The proposed approach demonstrated much robustness against the change of noise conditions compared with the original JA that has a single set of initial models.

The framework of Jacobian Adaptation can be extended further to the case where convolutive noise such as transfer channel changes as well as the additive noise dealt in this paper. The formulation for such complex environment will be presented in the next opportunity.

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*, volume 1, pages 233–236, 1992.
2. F. Martin, K. Shikano, , and Y. Minami. Recogion of noisy speech by composition of hidden Markov models. In *Proc. EuroSpeech-93*, pages 1031–1034, September 1993.
3. S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi. Jacobian approach to fast acoustic model adaptation. In *Proc. ICASSP*, pages 835–838, 1997.
4. S. Sagayama, Y. Yamaguchi, and S. Takahashi. Jacobian adaptation of noisy speech models. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (Santa Barbara)*, pages 396–403, 1997.
5. Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern. A Vector Taylor Series Approach for Environment-independent Speech Recognition. In *Proc. ICASSP*, volume 2, pages 733–736, 1996.
6. Y. Yamaguchi, J. Takahashi, S. Takahashi, and S. Sagayama. A Fast Acoustic Model Adaptation Technique Based on Taylor Series, September 1996. (in Japanese).