認識時に非観測な変動要因を考慮可能な音響モデリング

鈴木　浩之[†]　全　炳河[†]　南角　吉彦[†]　宮島千代美[††]　徳田　恵一[†]
北村　正[†]

† 名古屋工業大学　〒466-8555 愛知県名古屋市昭和区御器所町
†† 名古屋大学　〒466-8603 愛知県名古屋市千種区不老町
E-mail: †{h-suzuki,zen,nanaku,tokuda,kitamura}@ics.nitech.ac.jp, ††miyajima@is.nagoya-u.ac.jp

あらまし　本論文では，話者の声質や雑音など，認識時に非観測な変動要因を考慮した音声認識手法を提案する．連続音声認識における音響モデルとして，triphoneなどの前後の音素環境を考慮した音素環境依存モデルが広く利用されている．これらのモデルでは音響的な変動要因に対し音素を個別にモデル化することにより，モデルの精度を向上させることを目的にしている．本研究では，話者の声質や雑音なども音素の音響的な変動要因として考慮してモデルを構築する．このモデルを用いた認識では，triphoneにおける前後の音素や単語内の位置などが言語情報から与えられるのに対し，声質や雑音は認識時には未知であるため，何らかの方法でそれらの変動要因を決定する必要がある．この問題に対し，本研究では，認識時に非観測な変動要因を考慮したモデルを混合分布として統合することにより，入力音声が未知であっても認識可能な手法を提案する．音声認識実験の結果，提案法は従来法と比較して高い認識率が得られた．

キーワード　声質，雑音，音声認識，音響モデル，クラスタリング

Acoustic modeling in consideration of unknown variation factors at the time of recognition

Hiroyuki SUZUKI[†], Heiga ZEN[†], Yoshihiko NANKAKU[†], Chiyomi MIYAJIMA[††],
Keiichi TOKUDA[†], and Tadashi KITAMURA[†]

† Department of Computer Science and Engineering, Nagoya Institute of Technology　Gokiso-cho, Showa-ku, Nagoya, 466-8555, Japan
†† Department of Media Science, Nagoya University　Furo-cho, Chikusa-ku, Nagoya, 466-8603, Japan
E-mail: †{h-suzuki,zen,nanaku,tokuda,kitamura}@ics.nitech.ac.jp, ††miyajima@is.nagoya-u.ac.jp

Abstract　This paper proposes a speech recognition technique based on acoustic models which can take unknown variation factors (speaker voice characteristic, noise environment) into account at the time of recognition. Context-dependent acoustic models, which are typically triphone HMMs, are often used in continuous speech recognition systems. These methods enhance the accuracy of acoustic models via the respective modeling of phonemes according to the factors in acoustic variation. This work hypothesizes that the speaker voice characteristics that humans can perceive by listening and the noise environments are also factors of acoustic variation in construction of acoustic models, and a tree-based clustering technique is also applied to speaker voice characteristics and noise environments to construct proposed acoustic models. In speech recognition using triphone models, the neighboring phonetic context is given from the linguistic-phonetic knowledge in advance; in contrast, the variation factors as voice characteristics and noise environments of input speech are unknown in recognition using proposed acoustic models. This paper proposes a method of recognizing speech even under conditions where the variation factors of the input speech are unknown. The result of a gender-dependent speech recognition experiment shows that the proposed method achieves higher recognition performance in comparison to conventional methods.

Key words　voice characteristic, noise, speech recognition, acoustic model, clustering

1. INTRODUCTION

Phonetic-context-dependent acoustic models such as triphones that account for phonetic context before and after a phoneme are widely used in continuous speech recognition systems. The use of triphones rather than monophones is known to provide higher recognition accuracies. In speaker-independent speech recognition, variations in voice characteristics affect recognition performance. Furthermore, there are other factors of variation, e.g., environmental noise, speaking rate, and channel characteristics. As regards speaker voice characteristics, speaker adaptive training (SAT) [1] can be used to reduce the variability among speakers. However, speaker-independent models in the SAT system have to be adapted using adaptation data uttered by the target speaker beforehand.

This work proposes a simple technique for construction of unknown-variation-dependent models which can take into account variation factors unknown at recognition time. In this work, the speaker's voice characteristics and noise environments are assumed to be factors for variation that influence the acoustic characteristics of phonemes, unknown-variations-dependent acoustic models are constructed using a tree-based clustering technique based on MDL criteria [4].

Several methods of modeling the factors in phonetic variation such as the neighboring phonetic context as well as the position of a phoneme in a word and speaker's gender have been proposed [2], [3]. These methods enhance the accuracy of acoustic models via the respective modeling of phonemes according to the factors in acoustic variation. In contrast to the case of triphones where the neighboring phonetic context and phoneme position in a word are given by linguistic information, the voice characteristics or noise environments of the input speech have to be determined at the time of recognition in the proposed technique because they are unknown. This paper proposes a method for recognizing speech even under conditions where the types of voice characteristics and noise variations of the speech to input are unknown. Speech can be decoded using the proposed acoustic models with the constraint of fixing unknown variations in state, word, or sentence level. Since the voice characteristics are thought to rarely change in a sentence, sentence level decoding is the most rational approach. While, noise environment could be changed in an utterance. Therefore, the simplest approach without constraints is investigated in this work by integrating proposed acoustic models as a mixture distribution. This allows us to use a conventional speech decoder for recognition.

The rest of the paper is organized as follows. In Section 2, speech recognition using unknown-variations-dependent (proposed) acoustic models is described. Section 3 describes experimental results, and Section 4 notes conclusions and future topics.

2. Speech Recognition using Unknown-variations-dependent Acoustic Models

2.1 Constructing Unknown-variations-dependent Acoustic Models

To construct proposed acoustic models, training data must be labeled with regard to voice characteristics and noise environments. In this work, each speaker's voice characteristics are labeled according to the results of listening tests, and noise type and SNR are used as noise environments.

Since the set of possible phonetic-context-dependent models such as triphones for a standard language is very large, the estimation process often runs into the problem of a lack of data. To counter this problem, triphones must be grouped into a statistically estimable number of clusters and model parameters must be shared by using clustering methods such as tree-based context clustering [5]. In this work, a tree-based clustering technique is also applied to the speaker's voice characteristics or noise environments. Figures 1 and 2 show a binary decision tree that accounts for voice characteristics and noise environments as well as phonetic context, respectively. The questions about the voice characteristics or noise variations are used along with the conventional questions about the phonetic context for state cluster separation. The clustering is done for each state, and each cluster assumes a single Gaussian distribution. This simultaneous clustering of phonetic context and voice characteristics or noise environments enables the construction of proposed acoustic models that effectively share parameters.

2.2 Speech Recognition using Unknown-variations-dependent Acoustic Models

In speech recognition using triphone models, the adjacent phonetic context is given from the linguistic information in advance. In contrast, the voice characteristics or noise environments of the input speech are unknown in proposed acoustic models. A speech recognition method that allows recognition of speech even under conditions where types of unknown variations of the speech to input are unknown has been proposed.

In this method, leaf nodes having the same phonetic context and different voice characteristics or noise environments are integrated as a mixture distribution and the acoustic models are used in a conventional speech decoder. Figure 3 shows the integration method of the leaf nodes. With regard to questions about the phonetic context, either the "Yes" or "No" node is chosen as usual, and both "Yes" and "No" nodes are chosen with regard to questions about the

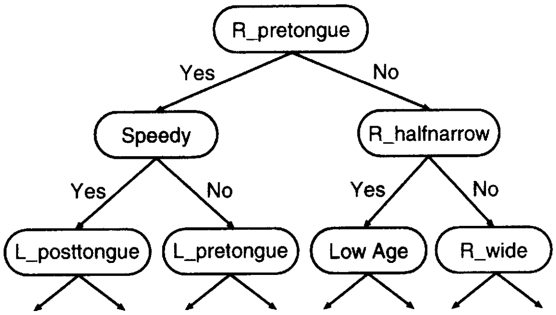


図1 声質を考慮した決定木に基づく状態クラスタリング.
Fig. 1 A decision tree considering voice characteristics.

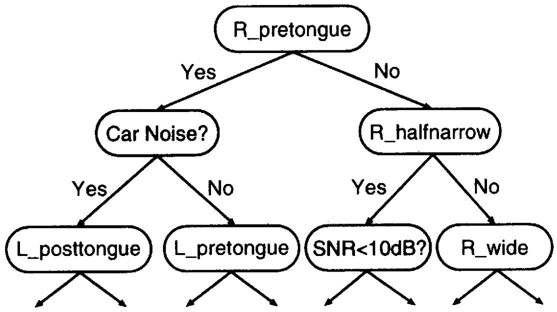


図2 雑音環境を考慮した決定木に基づく状態クラスタリング.
Fig. 2 A decision tree considering noise environments.

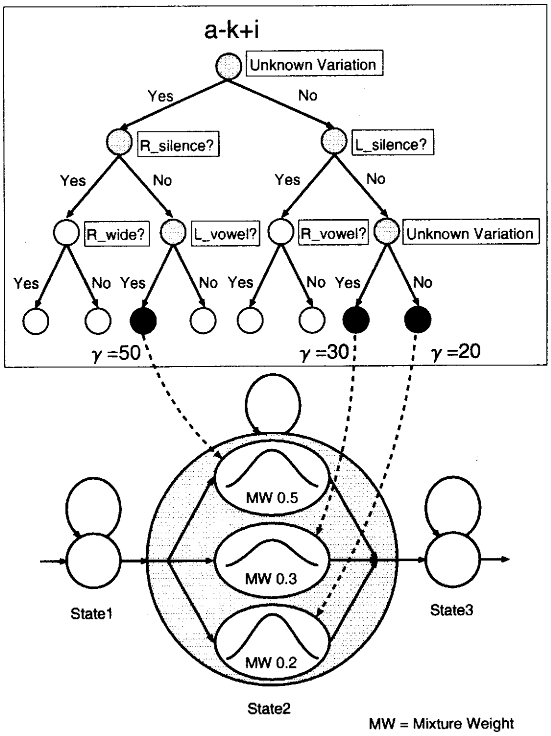


図3 音響モデルの統合.
Fig. 3 Integration of acoustic models.

unknown variations. By repeating these operations from the root node until reaching the leaf nodes, the set of leaf nodes that differs only in voice characteristics or noise environments is obtained for the respective phonetic-context-dependent triphones. The single Gaussian distributions of the leaf nodes are integrated as a new Gaussian mixture distribution, where the mixture weights are determined in proportion to the quantity of data γ (the accumulated state occupancy for the training data).

Through the aforementioned process, the integrated models can be used in the same speech decoder as for conventional triphone models. The voice characteristics of the input speech rarely change in a sentence. However, the noise environments often change in a sentence. Therefore, the simplest approach without the constraint of fixing voice characteristics or noise environments has been used in this experiment. Hence, frame-by-frame changes in the voice characteristics or noise environments are permitted in this recognition approach. From a different point of view, the integrated models are assumed to be independent of voice characteristics or noise environments. However, the more susceptible the triphone is to the difference in voice characteristics or noise environments, the more the mixture distributions are allocated to the triphone leading to efficient acoustic modeling with voice characteristics or noise environments taken into account.

3. Experimental Evaluation

In this paper, speech recognition experiment was done using voice-characteristic-dependent acoustic models. In the speech recognition under noise environment, modeling and recognition were performed in consideration of the information about noise. A model was then constructed simultaneously in consideration of voice characteristics and noise environments.

3.1 Databases

The ASJ-PB database (phonetically-balanced sentences) and ASJ-JNAS database (Japanese newspaper article sentences speech corpus) were used for training. About 20,000 sentences spoken by about 130 speakers of each gender were used for training. The IPA-98-TestSet that was not used for acoustic model training served as the test data. This test data consists of a total of 100 sentences spoken by 23 speakers of each gender.

In the experiment using noise environment as the variation factor, various noises recorded on the "Japan Electronic Industry Development Association noise database" were made to superimpose on the speech data, and study data and test data were created. Study data superimposed the noise of inside of running car (car), crossing, babble, and a factory. SNR are ∞, 20, 15, 10, and 5 dB to the speech data of clean. At this time, study data was divided into 17 sets and noise which is different in each was superimposed. Test data superimposed the noise of in a station concourse (station) and air-conditioning machine (aircon) in addition to four kinds of same noises as study data. SNR are ∞, 20, 10, and 0 dB to the speech data of clean.

3. 2 Experimental Conditions

To evaluate the proposed method, a gender-dependent speech recognition experiment was conducted. The speech data was windowed at a 10-ms frame rate using a 25-ms Blackman window, and parameterized into 12 mel-cepstral coefficients with a mel-cepstral analysis technique [6]. Static coefficients excluding zero-th coefficients and their first derivatives including zero-th coefficients were used as feature parameters. Cepstral mean subtraction was applied to each sentence. Three-state left-to-right HMMs were used to model 43 Japanese phonemes, and 146 phonological context questions and voice characteristic questions or noise environment questions were used to split nodes in decision trees.

MDL criteria [4] was used for context clustering, and in the proposed method, embedded training was applied before and after integrating proposed acoustic models. Under the aforementioned conditions, the proposed acoustic models were compared to conventional triphone models that take into account only the neighboring phonetic context.

3. 3 Labeling of Voice Characteristics

In this experiment using voice characteristics, 5 kinds of voice characteristic labels shown in Table 1 were used. A total of 40 listeners scored voice characteristics of the speakers used for training. Each characteristic label in Table 1 was scored by 4 different listeners with a 5-score ranking (from −2 to 2) and the score values of the 4 listeners were averaged and rounded off. Because of the difficulty in labeling all of the sentences for training, one randomly chosen sentence from the training data set of each speaker was presented to each listener. The speech data sets of males and females were listened to separately and labeled independently. Before each listening test, two voice samples that may have had the highest/lowest scores (2/−2) were presented to each listener so that the score distributions would not be biased.

3. 4 Results (voice characteristics)

As a result of clustering by the MDL criteria, the total number of distributions given by the proposed method was larger than that given by the conventional method. In order to compare the recognition performance with a comparable number of distributions, the number of mixtures of the conventional method was increased. The resulting total number of distributions are shown in Table 2. In the proposed method, it experimented by changing the number of voice characteristics to be used.

表 1 声質ラベルの詳細.

Table 1 Specification of voice characteristic label.

Label		Explanation of label
Age		Advanced / Low age
Cheerfulness		Cheerful / Dark
Sternness		Stern / Tender
Gender	Male	Masculine / Not masculine
	Female	Feminine / Not feminine
Speaking rate		Speedy / Slow

表 2 総分布数 (声質).

Table 2 Total number of distributions (voice characteristics).

	Conventional				Proposed		
	1-mix	2-mix	4-mix	8-mix	1-char	3-char	5-char
Male	7540	15080	30160	60320	8518	18719	32784
Female	7677	15354	30708	61416	8505	19614	33558

Figures 4 and 5 show the results of continuous speech recognition experiments for males and females, respectively. The results are shown in phoneme error rates. In the case of males, 1-char used only Cheerfulness, 3-char used Age, Cheerfulness, and Speaking rate, 5-char used all characteristics. When using 5 voice characteristics, slightly better performance was achieved when comparing the result of the proposed method to those of the conventional method of 4-mixture models. However, when using 1 or 3 voice characteristics, better result was obtained from the proposed method. In the cases of females, 1-char used only Sternness, 3-char used Age, Sternness, and Speaking rate, 5-char used all characteristics. In all patterns, better results were obtained from the proposed method than from the almost the same number of distributions of conventional method. The experimental results indicated that acoustic models of higher accuracy were constructed due to the efficient allocation of mixture distributions through modeling that accounted for voice characteristics.

3. 5 Consideration of noise environment

In the speech recognition experiment under noise environment, the kind and SNR of the superimposed noise were used as a label, although some labeling methods about the noise over study data were considered. As a label of SNR, each utterance unit (proposed1) and each phoneme unit (proposed2) were used. Furthermore, the label was given using both noise environments and voice characteristics. The details of labeling methods of SNR are shown in Table 3.

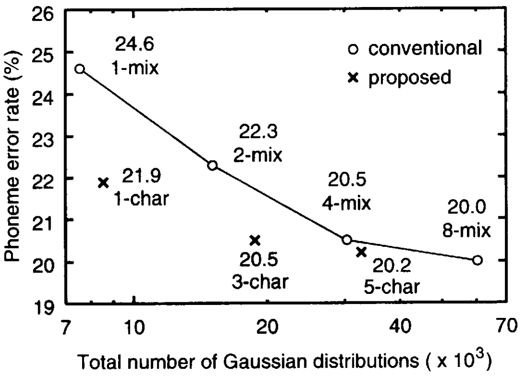


図 4 従来法と提案法の認識結果 (男性).

Fig. 4 Recognition results by the conventional method and the proposed method (male).

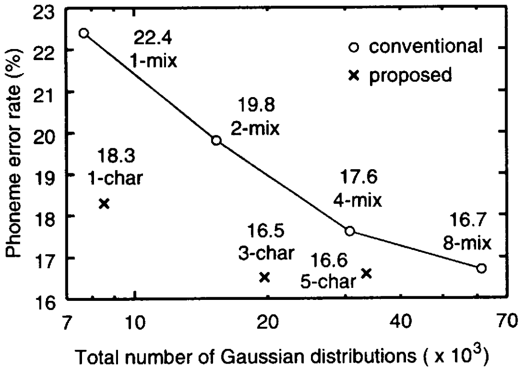


図 5 従来法と提案法の認識結果 (女性).

Fig. 5 Recognition results by the conventional method and the proposed method (female).

表 3 雑音のラベリング方法

Table 3 Labeling method of SNR

proposed1	The label of SNR is given per utterance
proposed2	The label of SNR is given in each phoneme unit (it quantizes per 3dB).
proposed3	proposed2 and a voice characteristic (male:Cheerfulness, female:Sternness)
proposed4	proposed2 and two voice characteristics (Cheerfulness and Sternness)

3.6 Results (noise environments)

As a result of clustering, the total number of distributions given by the proposed method was larger than that given by the conventional method. In order to compare the recognition performance with a comparable number of distributions, the number of mixtures of the conventional method was increased. The resulting total numbers of distributions are shown in Table 4.

Figures 6, 7, and Table 5 show the results of continuous speech recognition experiments for males and females, respectively. The SNR of the tests data of Figures 6, 7, and Table 5 are 0dB, 10dB, 20 dB, and clean, respectively. The number of distributions given by the proposed1 was almost the same as that given by the conventional method in the case of 2-mixture models. The number of distributions given by the proposed2, 3, and 4 was the almost middle of 1-mix and 2-mix of the conventional method. In the case of proposed1, almost same performance was achieved from the result when comparing the result of the conventional method of 4-mix, which is twice the total number of distributions. In the case of proposed2, almost same performance was achieved with the still fewer number of distributions than proposed1. The experimental results indicated that acoustic models of higher accuracy were constructed due to the efficient allocation of mixture distributions through modeling that accounted for noise environment.

表 4 総分布数 (雑音).

Table 4 Total number of distributions (noise).

	Conventional			Proposed			
	1-mix	2-mix	4-mix	1	2	3	4
Male	6457	12914	25828	11670	9228	9461	10277
Female	6421	12842	25684	12554	9835	10143	11241

表 5 従来法と提案法の認識結果 (クリーン).

Table 5 Recognition results by the conventional method and the proposed method (clean).

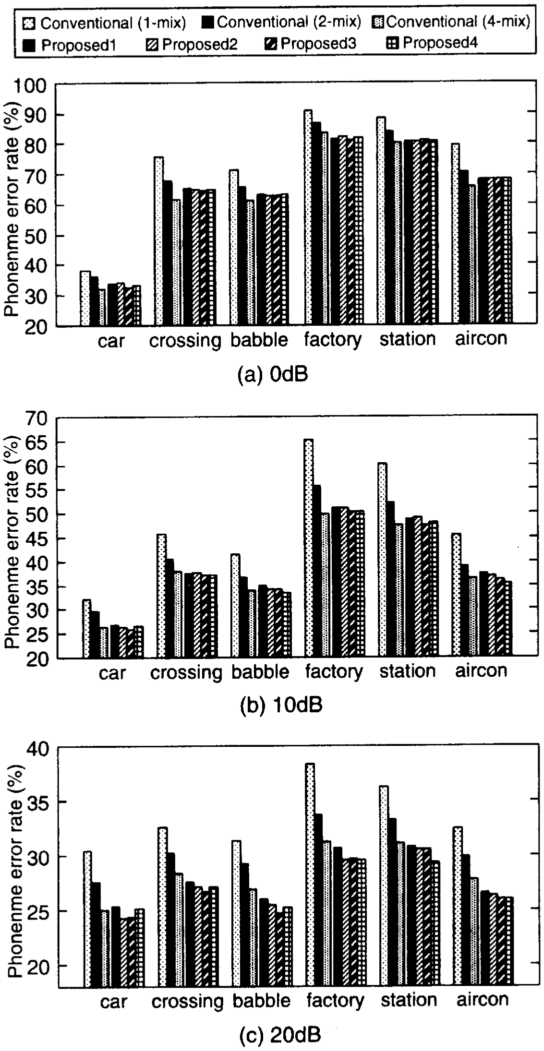
	Conventional			Proposed			
	1-mix	2-mix	4-mix	1	2	3	4
Male	33.7	29.3	26.8	25.9	25.8	25.5	25.4
Female	31.1	26.1	23.6	23.0	21.7	22.5	22.4

While, in proposed3 and 4 which also used voice characteristics as the variation factors simultaneously, improvement was not obtained for male and female. This shows not much that a speaker's voice characteristics does not affect it as the variation factors under a noise situation.

4. Conclusion

This paper has discussed the construction of voice-characteristic-dependent acoustic models or noise-environment-dependent acoustic models for speech recognition. The experimental results indicated that the proposed method outperformed the conventional method in terms of a comparable number of parameters.

As for future topics, the authors plan to conduct experiments with other speech decoding approaches using proposed acoustic models. The application of this method to large-vocabulary continuous speech recognition is also a future topic of interest.

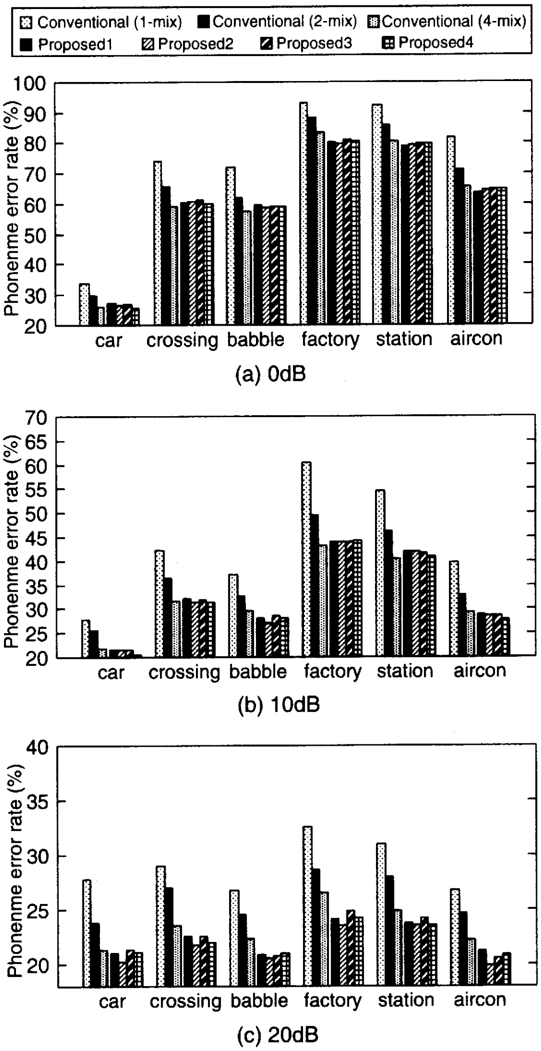


(a) 0dB

(b) 10dB

(c) 20dB

図 6 従来法と提案法の認識結果 (男性, S/N 比 = 0, 10, 20dB).
Fig. 6 Recognition results (male, SNR = 0, 10, 20dB).



(a) 0dB

(b) 10dB

(c) 20dB

図 7 従来法と提案法の認識結果 (女性, S/N 比 = 0, 10, 20dB).
Fig. 7 Recognition results (female, SNR = 0, 10, 20dB).

5. ACKNOWLEDGEMENTS

This work was partially supported by Grants for Researchers from the Hori Information Science Promotion Foundation and the Research Foundation for the Electro technology of Chubu, and a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science, Encouragement of Young Scientists (B) (Grant No. 14780274).

文　献

[1] T. Anastasakos, J. McDonough and J. Makhoul, “Speaker adaptive training: a maximum likelihood approach to speaker normalization,” *Proc. ICASSP'97*, pp. 1043–1046, 1997.

[2] W. Reichl and W. Chou, “A unified approach of incorporating general features in decision tree based acoustic modeling,” Proc. ICASSP'99, vol.2, pp.573–576, 1999.

[3] I. Shafran and M. Ostendorf, “Use of higher level linguistic structure in acoustic modeling for speech recognition,” Proc. ICASSP2000, vol.3, pp.1643–1646, 2000.

[4] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” J. Acoust. Soc. Jpn. (E), vol.21, no.2, pp.79–86, 2000.

[5] J. J. Odell, “The use of context in large vocabulary speech recognition,” PhD dissertation, Cambridge University, 1995.

[6] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” Proc. ICASSP'92, vol.1, pp.137–140, 1992.