SPEECH RECOGNITION USING VOICE-CHARACTERISTIC-DEPENDENT ACOUSTIC MODELS

H. Suzuki, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura

Department of Computer Science
Nagoya Institute of Technology
Nagoya, 466-8555, Japan

Email: {h-suzuki,zen,nankaku,chiyomi,tokuda,kitamura}@ics.nitech.ac.jp

ABSTRACT

This paper proposes a speech recognition technique based on acoustic models considering voice characteristic variations. Context-dependent acoustic models, which are typically triphone HMMs, are often used in continuous speech recognition systems. This work hypothesizes that the speaker voice characteristics that humans can perceive by listening are also factors in acoustic variation for construction of acoustic models, and a tree-based clustering technique is also applied to speaker voice characteristics to construct voice-characteristic-dependent acoustic models. In speech recognition using triphone models, the neighboring phonetic context is given from the linguistic-phonetic knowledge in advance; in contrast, the voice characteristics of input speech are unknown in recognition using voice-characteristic-dependent acoustic models. This paper proposes a method of recognizing speech even under conditions where the voice characteristics 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.

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. The use of triphones rather than monophones is known to provide higher recognition accuracy. In speech recognition using speaker-independent models, variations in voice characteristics affect recognition performance. 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 speaker-independent models considering the voice characteristic variations. There are various parameter clustering techniques for construction of context-dependent models [2, 3, 4]. In addition, 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 [5, 6]. These methods enhance the accuracy of acoustic models via the respective modeling of phonemes according to the factors in acoustic variation. The speaker's voice characteristics are also assumed to be factors for variation that influence the acoustic characteristics of phonemes and voice-characteristic-dependent acoustic models have been constructed using a tree-based clustering technique based on MDL criteria [7].

In speech recognition using voice-characteristic-dependent acoustic models, the voice characteristics of the input speech have to be determined because the voice characteristics are unknown at the time of recognition, in contrast to the case of triphones where the neighboring phonetic context and phoneme position in a word are given by linguistic information. This paper proposes a method of recognizing speech even under conditions where the types of voice characteristics of the speech to input are unknown. Speech can be decoded using the proposed acoustic models with the constraint of fixing voice characteristics 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. However, the simplest approach without constraints is investigated in this work by integrating voice-characteristic-dependent acoustic models as a mixture distribution. This allows use of a conventional speech decoder for recognition.

The rest of the paper is organized as follows. In Section 2, speech recognition using voice-characteristic-dependent

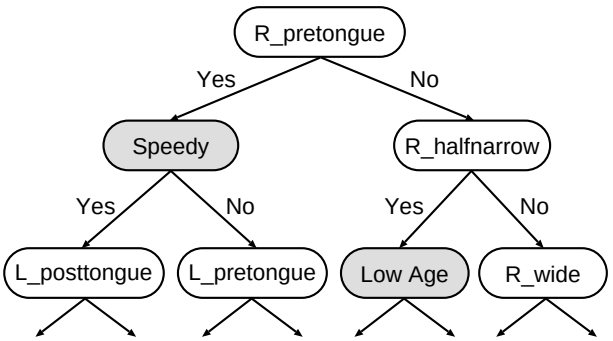

Fig. 1. A decision tree considering voice characteristics.

acoustic models is described. Section 3 describes experimental results, and Section 4 notes conclusions and future topics.

2. SPEECH RECOGNITION USING VOICE-CHARACTERISTIC-DEPENDENT ACOUSTIC MODELS

2.1. Constructing Voice-characteristic-dependent Acoustic Models

To construct voice-characteristic-dependent acoustic models, training data must be labeled with regard to voice characteristics. In this work, each speaker's voice characteristics are labeled according to the results of listening tests.

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 [8]. In this work, a tree-based clustering technique is also applied to the speaker's voice characteristics. Figure 1 shows a binary decision tree that accounts for voice characteristics as well as phonetic context. The questions about the voice characteristics 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 enables the construction of voice-characteristic-dependent acoustic models that effectively share parameters.

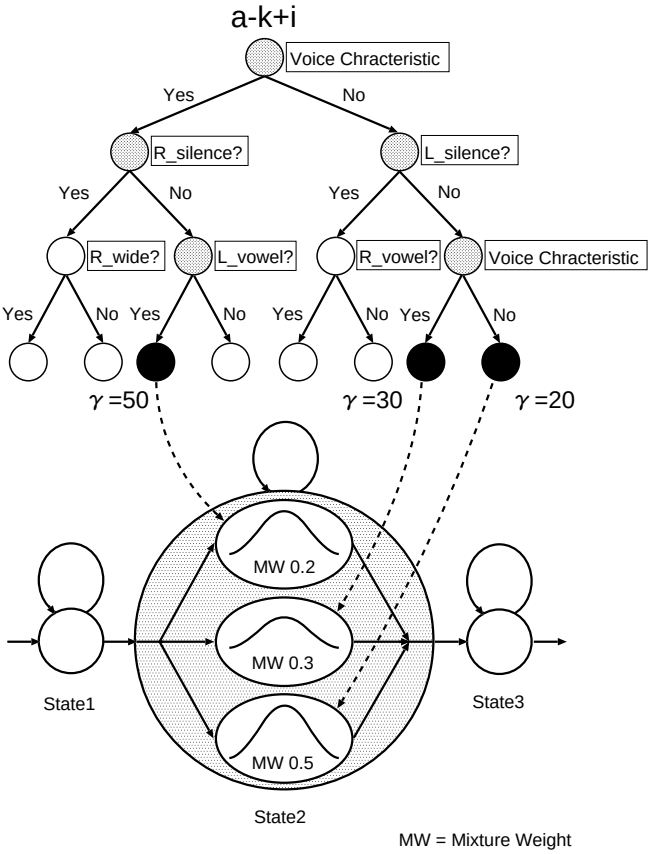

Fig. 2. Integration of voice characteristic dependent acoustic models.

2.2. Speech Recognition using Voice-characteristic-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 of the input speech are unknown in voice-characteristic-dependent acoustic models. A speech recognition method that allows recognition of speech even under conditions where types of voice characteristics of the speech to input are unknown has been proposed.

In this method, leaf nodes having the same phonetic context and different voice characteristics are integrated as a mixture distribution and the acoustic models are used in a conventional speech decoder. Figure 2 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 voice characteristics. By repeating these operations from the root node until reaching the leaf nodes, the set of leaf nodes that

differs only in voice characteristics 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 simplest approach without the constraint of fixing voice characteristics has been used in this experiment. Hence, frame-by-frame changes in the voice characteristics are permitted in this recognition approach. From a different point of view, the integrated models are assumed to be independent of voice characteristics. However, the more susceptible the triphone is to the difference in voice characteristics, the more the mixture distributions are allocated to the triphone leading to efficient acoustic modeling with voice characteristics taken into account.

3. EXPERIMENTAL EVALUATION

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.

3.2. Labeling of Voice Characteristics

In this experiment, 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.

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

Table 2. Total number of distributions (leaf-nodes).

	Proposed	Conventional			
		1-mix	2-mix	4-mix	8-mix
Male	32784	7540	15080	30160	60320
Female	33558	7677	15354	30708	61416

3.3. Experimental Conditions

To evaluate the proposed method, a gender-dependent speech recognition experiment was conducted. The speech data was down-sampled from 20kHz to 16kHz, 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 [9]. 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 20 voice characteristic questions were used to split nodes in decision trees.

MDL criteria [7] was used for context clustering, and in the proposed method, embedded training was applied before and after integrating voice-characteristic-dependent acoustic models. Under the aforementioned conditions, the voice-characteristic-dependent acoustic models (proposed method) were compared to conventional triphone models (conventional method) that take into account only the neighboring phonetic context.

3.4. Results

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 parameters, the number of mixtures of the conventional method was increased. The resulting total numbers of distributions are shown in Table 2. The

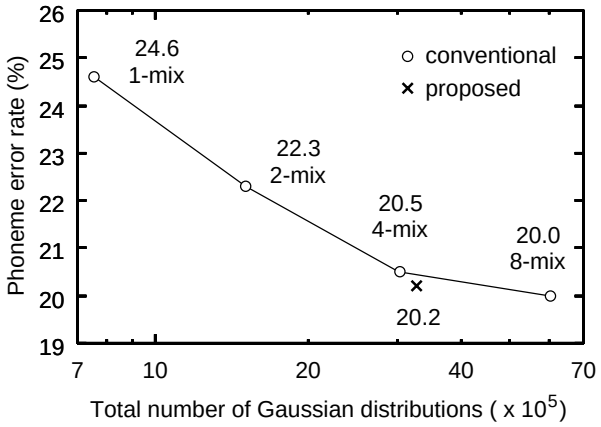

Fig. 3. Recognition results by the conventional method and the proposed method (male).

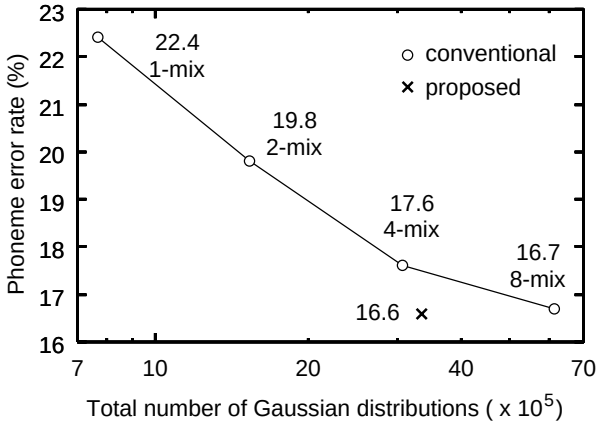

Fig. 4. Recognition results by the conventional method and the proposed method (female).

number of distributions given by the proposed method was almost the same as that given by the conventional method in the case of 4-mixture models.

Figures 3 and 4 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, slightly better performance was achieved when comparing the result of the proposed method to those of the conventional method of 4-mixture models, and in the cases of females, better results were obtained from the proposed method than from the conventional 8-mixture models. 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.

4. CONCLUSION

This paper has discussed the construction of voice-characteristic-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 voice-characteristic-dependent acoustic models. The application of this method to large-vocabulary continuous speech recognition is also a future topic of interest.

5. ACKNOWLEDGEMENTS

This work was partially supported by Grants for Researchers from the Hori Information Science Promotion Foundation and the Research Foundation for the Electrotechnology 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).

6. REFERENCES

[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] K.-F. Lee, "Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.38, no.4, pp.599–609, 1990.

[3] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," *Proc. ICASSP'92*, pp.573–576, 1992.

[4] M.-Y. Hwang, X. Huang, and F. Alleva, "Predicting unseen triphones with senones," *Proc. ICASSP'93*, pp.311–314, 1993.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.