PAPER *Special Section on Corpus-Based Speech Technologies*

Continuous Speech Recognition Based on General Factor Dependent Acoustic Models

Hiroyuki SUZUKI[†*], **Heiga ZEN**[†], ***Nonmembers***, **Yoshihiko NANKAKU**[†a)], **Chiyomi MIYAJIMA**[††], **Keiichi TOKUDA**[†], ***and*** **Tadashi KITAMURA**[†], ***Members***

SUMMARY This paper describes continuous speech recognition incorporating the additional complement information, e.g., voice characteristics, speaking styles, linguistic information and noise environment, into HMM-based acoustic modeling. In speech recognition systems, context-dependent HMMs, i.e., triphone, and the tree-based context clustering have commonly been used. Several attempts to utilize not only phonetic contexts, but additional complement information based on context (factor) dependent HMMs have been made in recent years. However, when the additional factors for testing data are unobserved, methods for obtaining factor labels is required before decoding. In this paper, we propose a model integration technique based on general factor dependent HMMs for decoding. The integrated HMMs can be used by a conventional decoder as standard triphone HMMs with Gaussian mixture densities. Moreover, by using the results of context clustering, the proposed method can determine an optimal number of mixture components for each state dependently of the degree of influence from additional factors. Phoneme recognition experiments using voice characteristic labels show significant improvements with a small number of model parameters, and a 19.3% error reduction was obtained in noise environment experiments.

key words: *continuous speech recognition, triphone HMMs, context clustering, Bayesian networks, voice characteristic, noise environment*

1. Introduction

In recent large vocabulary speech recognition systems, context-dependent acoustic models, i.e., triphone HMMs, are commonly used. Since triphone HMMs are trained individually depending on the phoneme contexts, the acoustic features are modeled more accurately than the monophone HMMs. However, in the use of context-dependent HMMs, there is the problem that the number of training data becomes relatively small for each triphone HMM due to the large number of triphones and the statistical reliability of HMMs is reduced. To adjust the balance between the model complexity and the number of training data, various parameter sharing techniques have been proposed [1]–[5]. Especially, the tree-based context clustering [5] is well employed with triphone HMMs. It has two advantages over bottom-up based approaches: first, by incorporating phonetic knowledge into questions, it can assign unseen context-dependent HMMs to the leaf nodes of decision trees. Second, the splitting procedure of the decision tree provides a way of keeping the balance of model complexity and robustness.

In recent years, several attempts to utilize not only phonetic contexts but additional contexts (factors) which vary the distribution of acoustic features have been made in HMM-based acoustic modeling, e.g., phoneme position in a word [6], speaker's gender [7], and dialect, speaking rate and SNR [8]. Although accurate modeling of acoustic feature can be performed by using explicit information of factors, when the additional factors for testing data are unobserved, methods for obtaining factor labels is required independently of triphone HMMs. In [9], incorporating the complemental features into the HMM-based acoustic modeling have been described within a framework of Dynamic Bayesian Networks (DBNs). In this method, the Gaussian mixture density function is replaced with the Bayesian Networks (BNs) as a state output distribution, and the additional factors can be included as a discrete probabilistic variables into BNs. The derived equations from simple BN structure is interpreted as that the likelihood of hybrid HMM/BN is calculated by summing over the likelihoods of the factor dependent HMMs. In the recognition phase, the unobserved factors are dealt as hidden variables and marginalized over the possible values of factors.

In this paper, we propose a model integration technique for decoding speech based on the general factor dependent HMMs. The integrated models over additional factors can be regarded as factor invariant models based on the framework of Bayesian networks. By ignoring the temporal dependency between additional factors, the integrated models can be used by the conventional decoder as a standard triphone HMM with Gaussian mixture density functions. Although the proposed method is similar with the hybrid HMM/BN, the proposed method uses the results of context clustering based on the MDL (Minimum Description Length) criterion [10], and the optimal number of mixture components for each state can be determined dependently of the degree of influence from additional factors.

The rest of the paper is organized as follows. In Sect. 2, tree-based context clustering for additional factor dependencies and a model integration technique for decoding is described. Section 3 presents experimental results of the proposed method in which voice characteristics and noise-environment were adopted as the additional contexts. Finally, Sect. 4 notes conclusions and future topics.

Manuscript received July 8, 2004.

Manuscript revised September 26, 2004.

[†] The authors are with the Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.

[††] The author is with the Department of Media Science, Nagoya University, Nagoya-shi, 466–8603 Japan.

[*] Presently, with DENSO Corporation.

a) E-mail: nankaku@ics.nitech.ac.jp

DOI: 10.1093/ietisy/e88–d.3.410

2. General Factor Dependent Acoustic Modeling

2.1 Tree-Based Context Clustering for General Factors

In this paper, we consider not only phonetic contexts but any additional factors which vary the distributions of acoustic features, in order to make context dependent HMMs more accurate. In general factor dependent acoustic models, training data is divided by all the combination of triphone contexts and additional factors, then modeled individually. Although accurate modeling of acoustic feature can be performed by using explicit information of factors, the number of training data for each HMM is decreased due to a number of factor dependent HMMs. Thus, the reliability of factor dependent HMMs is reduced. Furthermore there also exist many triphones which are not observed in training data so-called unseen triphones. However, similar to the standard triphone HMMs, this problem can be avoided by applying the tree-based context clustering technique.

In the tree-based context clustering, triphone HMMs are grouped into "clusters," and all HMMs belonging to one cluster are assumed to have the same model parameters. A binary tree is constructed based on the maximum likelihood criterion by applying a question of "Yes" or "No" to each node and splitting the cluster into two child clusters iteratively. By limiting the number of possible splitting using prior knowledge, linguistic and articulatory information can be reflected in the clustering results. The procedure of the tree-based context clustering is as follows:

Step 1: Create a root node which includes all triphone HMMs and compute its likelihood.

Step 2: The likelihood when a question would be applied is calculated for all questions in each leaf node.

Step 3: Select the pair of node and question which gives the maximum likelihood, and split it into two by applying the question.

Step 4: If the change of the likelihood after splitting is below a threshold, stop the procedure. Otherwise, go to Step 2.

After the procedure, the clusters represented by leaf nodes are used as the sharing structure, and any triphone HMM including unseen triphones are assigned to one of clusters. A cluster to which a target triphone HMMs belongs can be found by descending the constructed tree from the root to the leaf node while answering the questions at each node based on the target label.

We can simply apply the tree based clustering technique to the general factor dependent HMMs by preparing the questions about additional factors. Figure 1 shows an example of binary decision tree for general factor-dependent HMMs. In the figure, the white and gray nodes represent the nodes with questions about a phonetic context and an additional factor, respectively. This simultaneous clustering of phonetic contexts and additional factors enables effective sharing parameters of acoustic models. In this paper, we assumed that the clustering is performed at each state-position, and each cluster is modeled as a single Gaussian distribution. Instead of the maximum likelihood criterion, the Minimum Description Length (MDL) criterion is adopted, in which the optimal number of clusters is determined automatically without setting a threshold. The description length can be calculated as follows:

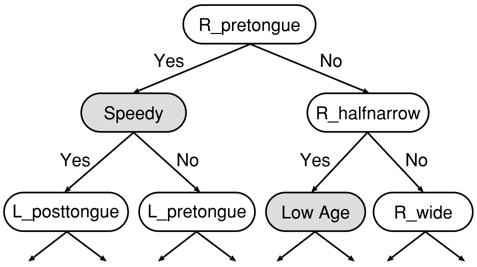


Fig. 1 A decision tree considering additional contexts.

$$DL = \frac{1}{2}\sum_{m=1}^{M}\Gamma_m(K + K\log(2\pi) + \log|\boldsymbol{\Sigma}_m|) \\ +cKM\log\Gamma_0 + const \quad (1)$$

where Γ_m and $\boldsymbol{\Sigma}_m$ are the total state occupancy count and the covariance matrix of the leaf node $m = 1, \ldots, M$, respectively. The total number of data is denoted by $\Gamma_0 = \sum_{m=1}^{M}\Gamma_m$ and K is the dimensionality of the feature vectors. Although in [10], a weight coefficient c is adjusted to control the size of decision trees, we simply use $c = 1.0$ in our experiments.

2.2 Model Integration Technique for Decoding

When input speech is decoded by standard triphone HMMs, the adjacent phonetic contexts can be given by the phonetic connections in search of a network. However, some additional factors, such as voice characteristics, speaking rate, and SNR, etc., cannot be determined without an estimation technique which is independent of the acoustic modeling. If we have a reliable method for obtaining the explicit values of additional factors, it would be able to improve the recognition performance by using factors as observed variables. However, there is no such method necessarily for additional factors; the factors should be dealt as hidden variables, and integrated out based on the framework of the Bayesian networks. That is, the likelihood function for an HMM Λ with hidden factor variables $\boldsymbol{c}$ can be written as:

$$\begin{aligned} P(\boldsymbol{O}|\Lambda) &= \sum_{\boldsymbol{q}}\sum_{\boldsymbol{c}} P(\boldsymbol{q}|\Lambda)P(\boldsymbol{O},\boldsymbol{c}|\boldsymbol{q},\Lambda) \\ &= \sum_{\boldsymbol{q}} P(\boldsymbol{q}|\Lambda)\left\{\sum_{\boldsymbol{c}} P(\boldsymbol{O},\boldsymbol{c}|\boldsymbol{q},\Lambda)\right\} \quad (2) \end{aligned}$$

where $\boldsymbol{O} = (\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T)$ and $\boldsymbol{q} = (q_1, \ldots, q_T)$ are the observed data and the state sequence of an HMM, respectively. The notation $\boldsymbol{c} = (c_1, \ldots, c_T)$ means the additional factors. By ignoring the temporal dependency between additional factors, the output probability of each time can be calculated individually as a Gaussian mixture density:

$$\begin{aligned}\sum_{\boldsymbol{c}} P(\boldsymbol{O}, \boldsymbol{c}|\boldsymbol{q}, \Lambda) &= \sum_{\boldsymbol{c}} P(\boldsymbol{c}|\boldsymbol{q}, \Lambda)P(\boldsymbol{O}|\boldsymbol{c}, \boldsymbol{q}, \Lambda)\\ &\simeq \sum_{\boldsymbol{c}} \prod_{t=1}^{T} P(c_t|q_t, \Lambda)P(\boldsymbol{o}_t|c_t, q_t, \Lambda)\\ &= \prod_{t=1}^{T} \left\{ \sum_{c_t} P(c_t|q_t, \Lambda)P(\boldsymbol{o}_t|c_t, q_t, \Lambda) \right\} \qquad (3)\end{aligned}$$

where $P(\boldsymbol{o}_t|c_t, q_t, \Lambda)$ is the factor dependent Gaussian distribution and $P(c_t|q_t, \Lambda)$ is the weight of mixture components. If the additional factors c_t are unsupervised in the model training, the integrated models is equivalent to the standard HMMs with multi-mixture Gaussian density functions.

In the proposed technique, the leaf nodes of the decision tree which have the same phonetic contexts and different additional factors are integrated as a mixture of Gaussian distribution. The integration technique of the leaf nodes is shown in Fig. 2 and the procedure is summarized as follows:

Step 1: If the question of a current node is about phonetic contexts, either "Yes" or "No" node is chosen. Otherwise, i.e., the question is about an additional factor, both "Yes" and "No" nodes are chosen.

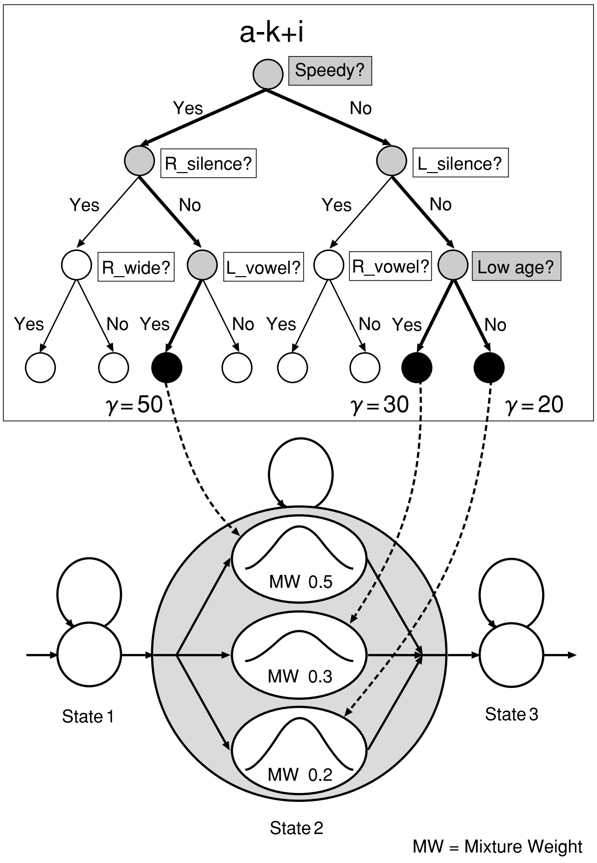


Fig. 2 Integration of acoustic models.

Step 2: By repeating Step 1 from the root node until reaching all leaf nodes, a set of leaf nodes can be obtained, which has the same triphone context but different additional factors.

Step 3: The single Gaussian distributions of the leaf nodes $P(\boldsymbol{o}_t|c_t, q_t, \Lambda)$ are integrated as a new mixture distributions. The mixture weights $P(c_t|q_t, \Lambda)$ are determined in proportion to the quantity of data γ (the accumulated state occupancy count for the cluster).

Through the above procedure, the integrated models can be obtained from the factor dependent HMMs. The integrated model have the same structure as the standard HMMs except that a Gaussian distribution in each state corresponds to a cluster of the additional factors. Therefore, the integrated models can be used by the conventional decoders of triphone HMMs without dictionary conversion and/or decoder modification. Furthermore, the constructed HMMs have an optimal number of mixtures in each state, which are determined by the degree of influence from the additional factors.

3. Experiments

To evaluate the performance of the proposed method, continuous speech recognition experiments were conducted using voice characteristics and noise environment as the additional factors.

3.1 Experimental Conditions

The ASJ-PB database (phonetically-balanced sentences) and ASJ-JNAS database (Japanese newspaper article sentences speech corpus) were used. About 20,000 sentences spoken by about 130 speakers of each gender were used for training. For testing, the IPA-98-TestSet was used, which consists of a total of 100 sentences spoken by 23 speakers for each gender. In the experiment using noise environment, the Japan Electronic Industry Development Association (JEIDA) noise database was used, and dissimilar 6 kinds of noise data were chosen from the total 17 kinds of noise data in the database. The training speech data was divided into 17 sets, and one was used for clean speech, and the other were used for noise environment. The noise data of "inside of running car (car),""crossing,""babble," and "factory" with SNR of 20, 15, 10, and 5 dB were superimposed on the each set of clean speech data. For testing data, in addition to the four kinds of noise for training, noise data of "in a station concourse (station)" and "air-conditioning machine (aircon)" were superimposed with SNR of 20, 10, and 0 dB.

The speech data was sampled to 16 kHz, 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 [11]. The static coefficients excluding zero-th coefficients and their first derivatives including zero-th coefficients were used as feature vectors. The

CMS (cepstral mean subtraction) was applied to each utterance. These acoustic features were modeled by 3 states left-to-right HMMs of 43 Japanese phonemes. Continuous phoneme recognition experiments were performed. A network to limit the results to Japanese phoneme sequences was used. In the tree-based context clustering, 146 phonological context questions were prepared for conventional triphone HMMs. In addition, 20 voice characteristic questions and 43 noise environment questions were prepared for the additional factors. For both the conventional and the proposed method, the MDL criterion was used as the stopping rule for the context clustering. In the proposed technique, the embedded training was conducted before and after the model integration.

3.2 Labeling Methods

To construct context-dependent HMMs which depend on the additional factors, training data was labeled with regard to voice characteristics and noise environments. To select the kinds of labels for the proposed method, it is necessary to consider two properties: the additional factors should have strong dependency on the acoustic features, and should be independent of each other with respect to acoustic features. However, it is difficult to select the labels which satisfy the above properties in practice, hence the voice characteristic labels in this experiment were determined so as to be easily scored in listening test. Table 1 shows the kinds of voice characteristic labels used in this experiment. A total of 40 listeners scored voice characteristics of the training data. Because of the large number of training data, we assumed that speech data uttered by one speaker have the same characteristics, so that the labels scored from one sentence (randomly chosen) were used as those of all training data for the corresponding speaker. Each characteristic was scored on 5-levels by four listeners and the average of the four listeners was rounded off and used as the labels. Before each listening test, two voice samples that may have had the highest/lowest scores were presented to each listener so that the score distributions would not be biased.

For noise environment, the noise kinds and SNR were used as the labels. The values of SNR were quantized at 3 dB intervals. We also prepared two sets of labels: SNR is calculated for each utterance (noise1) and each phoneme (noise2).

3.3 Experimental Results

Table 2 presents the number of Gaussian distributions after the context clustering based on voice characteristic dependent HMMs. Using the MDL criterion, the total number of distributions obtained by the proposed method become larger than that of the conventional HMMs without additional factors. To evenly compare the recognition performance with a similar number of model parameters, the number of Gaussian distributions was increased to 2, 4 and 8 mixtures for all states of the conventional triphone HMMs.

Table 1 Voice characteristic labels.

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 (voice characteristics).

techniques		Male	Female
Conventional	1-mix	7540	7677
	2-mix	15080	15354
	4-mix	30160	30708
	8-mix	60320	61416
Proposed	1-char	8076	8522
	2-char	12442	13399
	3-char	19213	20665
	4-char	26602	28116
	5-char	32784	33558

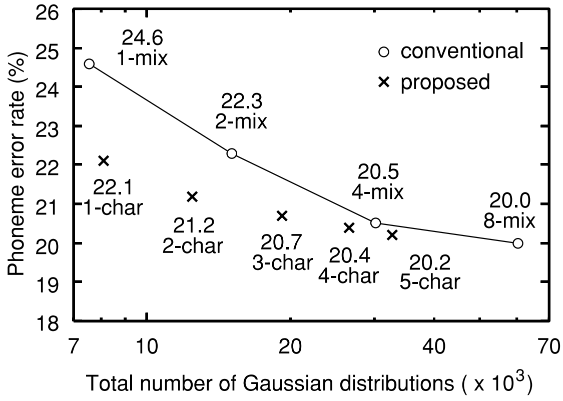


Fig. 3 Recognition results with voice characteristics labels (male).

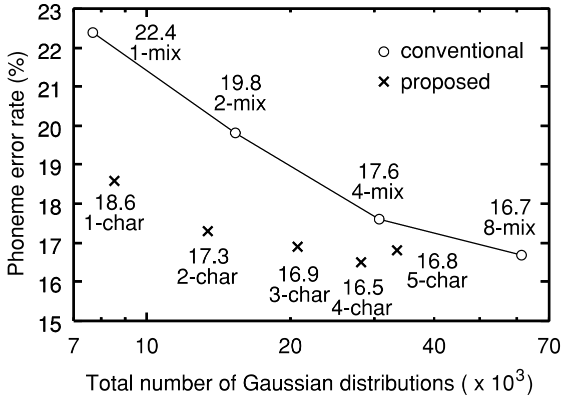


Fig. 4 Recognition results with voice characteristics labels (female).

The number of kinds of voice characteristics incorporated into the triphone HMMs was also changed, and the average error rate of all combinations of the voice characteristics was used as the results. Figures 3 and 4 show the phoneme error rates of continuous speech recognition for males and females, respectively. In the case of "5-char," the proposed method achieved slightly better performance than the conventional HMMs with 4-mixture which have almost the same number of parameters. However, the error rate of "5-char" was higher than "4-char" in the female case, and "5-char" could not outperform "8-mix" in both the male and female case. This might be because of inaccuracy of the voice characteristic labels obtained by the listening test. The choice of label kinds is also a possible reason, that is, there still exist other dominant factors with respect to acoustic features. However, when using 1, 2, 3, and 4 voice characteristics, a significant error reduction was achieved by the model integration without using labels of voice characteristics for testing data. In the recognition phase, calculating the posterior probabilities of Gaussian mixtures corresponds to the estimation of voice characteristics. According to the recognition results, it can be considered that appropriate Gaussian components would be selected probabilistically in the integrated HMMs. Furthermore, the integrated HMMs have an optimal number of Gaussians for each state dependently of the degree of influence from voice characteristics, this would lead to the improvement in the cases of small number of model parameters.

In the experiments with noise superimposed data, the total number of distributions obtained by the context clustering are shown in Table 3. The number of distributions given by the noise dependent HMMs was also larger than the conventional HMMs with single Gaussian distributions. This is because the Gaussian distributions were split by the questions dependently of the noise kinds and SNR in the context clustering. Table 4 shows the phoneme error rates of clean test data, and Figs. 5 and 6 present the results of noise superimposed test data with SNR of 0 dB, 10 dB and 20 dB. In the case of clean speech data, even though the number of distributions is smaller than the 4-mixture case, the proposed method outperformed the conventional HMMs without using labels of noise kinds and its SNR for testing data. Furthermore, for all kinds of noise with 20 dB, the proposed method still achieved higher recognition rates than the 4-mix of the conventional triphone HMMs, even though the integrated HMMs have only half the number of model parameters of the 4-mix models. An error reduction rate of 19.3% was obtained by "noise2" of female, 20 dB and crossing noise (21.8%) as compared with "2-mix" (27.0%) which have the similar number of parameters. In the Figs. 5 and 6, the effectiveness of the proposed method was reduced with decreasing SNR, and in the case of 0 dB, no significant improvement was found by using the model integration method. However, the proposed method could obtain slightly better performance as compared to 2-mix which has similar number of distributions.

Note that the results with noises which did not appear in the training data ("station" and "aircon") are also superior than the conventional method in the 20 dB case. The noise kind and SNR of testing data are estimated as the posterior probability distribution of the Gaussians corresponding to the noise and SNR of the training data. Therefore it can be

Table 3 Total number of distributions (noise environments).

techniques		Male	Female
Conventional	1-mix	6457	6421
	2-mix	12914	12842
	4-mix	25828	25684
Proposed	noise1	11670	12554
	noise2	9228	9835

Table 4 Recognition results of clean speech data with noise environments labels.

techniques		Male	Female
Conventional	1-mix	33.7	31.1
	2-mix	29.3	26.1
	4-mix	26.8	23.6
Proposed	noise1	25.9	23.0
	noise2	25.8	21.7

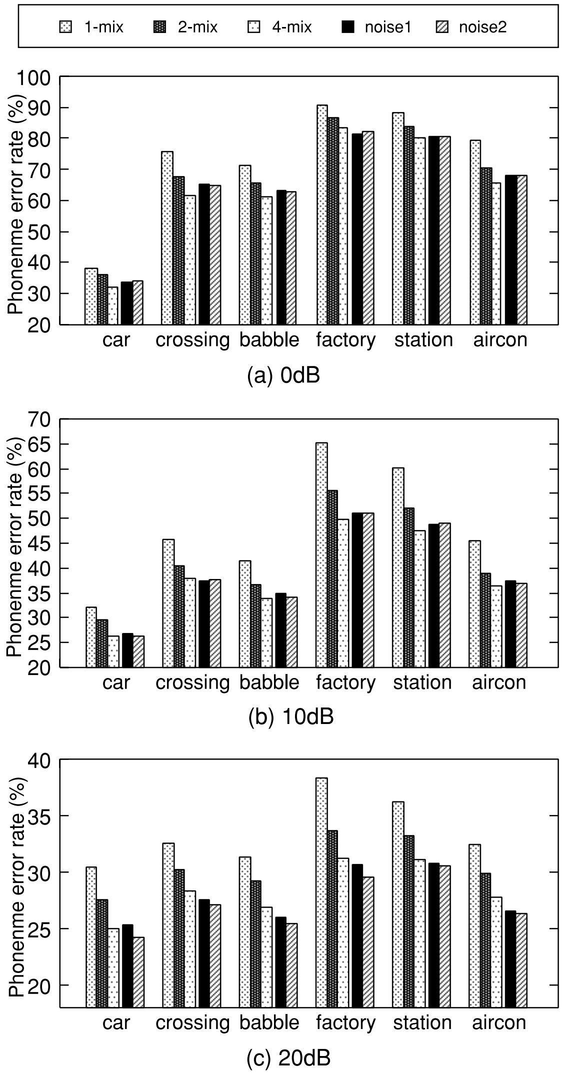


Fig. 5 Recognition results with noise (male). "1-mix," "2-mix" and "4-mix" mean the conventional HMMs with increasing the number of Gaussians. "noise1" and "noise2" are the integrated HMMs which were trained with SNR labels of utterance-level and phoneme-level, respectively.

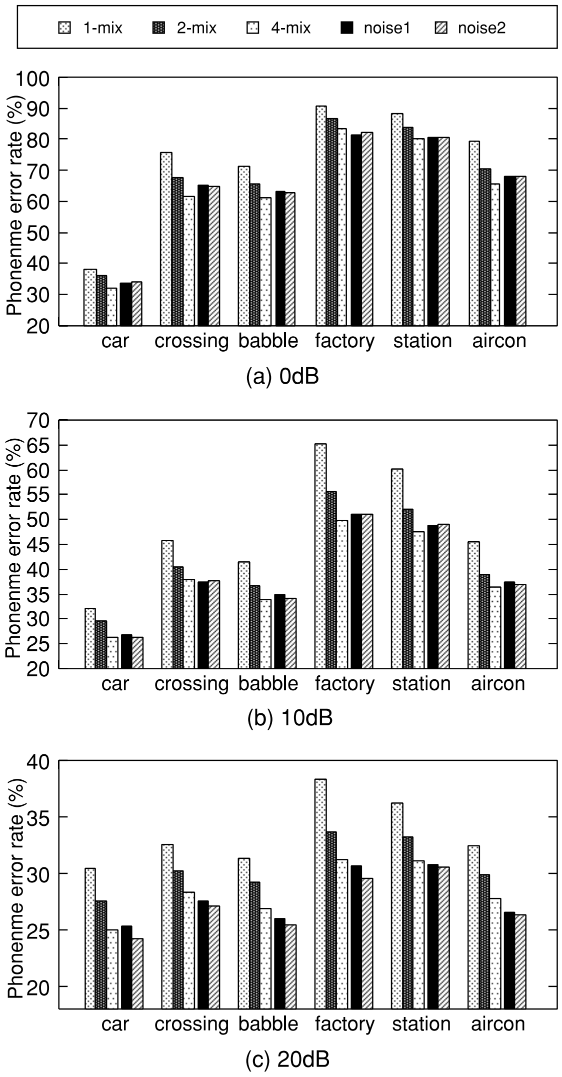


Fig. 6 Recognition results with noise (female). "1-mix," "2-mix" and "4-mix" mean the conventional HMMs with increasing the number of Gaussians. "noise1" and "noise2" are the integrated HMMs which were trained with SNR labels of utterance-level and phoneme-level, respectively.

considered that the training data in this experiment has sufficient variation to represent the acoustic features of the open noise data. This result would also be caused by a smoothing effect of the noise kinds and SNR, which was performed by the embedded training after the context clustering. Although the integrated HMMs after the embedded training are equivalent to the standard HMMs with Gaussian mixture densities, because of the initialization dependence problem of the EM algorithm, the different model parameters were obtained by the proposed method; the factor dependent models give a good initial point for the EM algorithm based on complemental information of additional factors. Moreover, the proposed method has the advantage that the optimal number of Gaussians can be determined for each state by using the results of tree-based context clustering.

From the results of the voice characteristics and noise environment, it is confirmed that the proposed method can model acoustic features efficiently based on the factor dependent HMMs and the context clustering. Since the integrated HMMs can be used by the conventional decoder without using the labels of additional factors, the proposed method can simply be applied to the other kinds of additional factors which complement acoustic features.

4. Conclusion

This paper has described the framework of continuous speech recognition using general factor dependent acoustic models.

The proposed model integration technique constructs a factor invariant models which can be used by the conventional decoders without dictionary conversion and/or decoder modification. Since the integrated model ignores the correlation between additional factors of consecutive frames, the model structure is similar to the standard HMMs with Gaussian mixture densities. However, the proposed method has an advantage that the factor dependent models give good initial parameters in the embedded training for the integrated HMMs. Consequently the EM algorithm can estimate model parameters which well represent acoustic variations of training data by hidden variables corresponding to the additional factors. Furthermore, by using the MDL based context clustering, the proposed method can determine an optimal number of mixture components for each state, dependently of the degree of influence from the additional factors. This efficient modeling improves recognition performance especially when the total number of model parameters is relatively small.

The experiments using the voice characteristics labels showed that the proposed model achieved similar performance as conventional triphone HMMs with only less than half number of model parameters. In the noise environment experiments, a 19.3% error redocution was obtained in the case of female, 20 dB and crossing noise. Moreover, because of the smoothing effect by the embedded training, the results with open noise data are also superior than the conventional method in the 20 dB case.

The authors plan to conduct experiments with decoding approaches using temporary correlation of factors. Investigation of different choices of voice characteristic labels and effectiveness to other kind of noises are also future works.

Acknowledgments

This paper 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).

References

[1] K.F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," IEEE Trans. Acoust., Speech Signal Process., vol.38, no.4, pp.599–609, 1990.

[2] P.C. Woodland and S.J. Young, "Benchmark DARPA RM results with the HTK portable HMM toolkit," Proc. DARPA Continuous Speech Recognition Workshop, pp.71–76, 1992.

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

[4] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," Proc. ICASSP, vol.I, pp.573–576, 1992.

[5] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," Proc. ICASSP, pp.307–311, 1994.

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

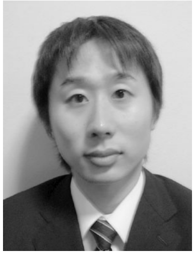
[7] I. Shafran and M. Ostendorf, "Use of higher level linguistic structure in acoustic modeling for speech recognition," Proc. ICASSP, vol.2, pp.1021–1024, 2000.

[8] C. Fugen and I. Rogina, "Integrating dynamic speech modalities into context decision trees," Proc. ICASSP, vol.III, pp.1277–1280, 2000.

[9] K. Markov and S. Nakamura, "A hybrid HMM/BN acoustic model for automatic speech recognition," IEICE Trans. Inf. & Syst., vol.E86-D, no.3, pp.438–445, March 2003.

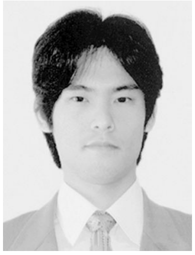
[10] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), vol.21, no.2, pp.79–86, 2000.

[11] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. ICASSP, vol.1, pp.137–140, 1992.



Hiroyuki Suzuki received the B.E. degree in computer science and M.E. degree in electrical and computer engineering from Nagoya Institute of Technology, Nagoya, Japan, in 2002 and 2004, respectively. He is currently with DENSO Corporation.

Heiga Zen received the A.E. degree in electronic and information engineering from Suzuka National College of Technology, Suzuka, Japan, and B.E. degree in computer science and M.E. degree in electrical and computer engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1999, 2001 and 2003, respectively. Currently he is a Ph.D student of the Department of Computer Science at the Nagoya Institute of Technology. His research interests include automatic speech recognition and text-to-speech synthesis. He is a member of ASJ.



Yoshihiko Nankaku received the B.E. degree in Computer Science, and the M.E. and Dr.Eng. degrees in the Department of Electrical and Electronic Engineering from the Nagoya Institute of Technology, Nagoya Japan, in 1999, 2001, and 2004 respectively. He is currently a postdoctoral fellow at the Nagoya Institute of Technology. His research interests include statistical machine learning, image recognition, speech recognition and synthesis and multimodal interface. He is a member of ASJ.

Chiyomi Miyajima received the B.E. degree in computer science and M.E. and Dr.Eng. degrees in electrical and computer engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1996, 1998, and 2001, respectively. From 2001 to 2003, she was a Research Associate of the Department of Computer Science, Nagoya Institute of Technology. Currently she is a Research Associate of the Graduate School of Information Science, Nagoya University, Nagoya, Japan. Her research interests include automatic speaker recognition and multi-modal speech processing. She received the Awaya Award in 2000 from the Acoustical Society of Japan. She is a member of ASJ and JASL.

Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004 he was a Associate Professor at the Department of Computer Science, Nagoya Institute of Technology as Associate Professor, and now he is a Professor at the same institute. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories, Japan and was a Visiting Researcher at Carnegie Mellon University from 2001 to 2002. He is a co-recipient of the Paper Award and the Inose Award both from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. His research interests include speech speech coding, speech synthesis and recognition, and statistical machine learning.

Tadashi Kitamura received the B.E. degree in electronics engineering from Nagoya Institute of Technology, Nagoya, in 1973, and M.E. and Dr. Eng. degrees from Tokyo Institute of Technology, Tokyo, in 1975 and 1978, respectively. In 1978 He joined the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology, as a Research Associate. In 1983 he joined Nagoya Institute of Technology, as a Assistant Professor. He is currently a Professor of Graduate School of Engineering of Nagoya Institute of Technology. His current interests include speech processing, image processing and multi-modal biometrics. He is a member of IEEE, ISCA, ASJ, IPSJ and ITE.