IEICE TRANS. INF. & SYST., VOL.E93–D, NO.9 SEPTEMBER 2010

PAPER Special Section on Processing Natural Speech Variability for Improved Verbal Human-Computer Interaction

Learning Speech Variability in Discriminative Acoustic Model Adaptation

Shoei SATO[†a)], **Takahiro OKU**[†], **Shinichi HOMMA**[†], ***Members***, **Akio KOBAYASHI**[†], ***Nonmember***, ***and*** **Toru IMAI**[†], ***Member***

SUMMARY We present a new discriminative method of acoustic model adaptation that deals with a task-dependent speech variability. We have focused on differences of expressions or speaking styles between tasks and set the objective of this method as improving the recognition accuracy of indistinctly pronounced phrases dependent on a speaking style.The adaptation appends subword models for frequently observable variants of subwords in the task. To find the task-dependent variants, low-confidence words are statistically selected from words with higher frequency in the task's adaptation data by using their word lattices. HMM parameters of subword models dependent on the words are discriminatively trained by using linear transforms with a minimum phoneme error (MPE) criterion. For the MPE training, subword accuracy discriminating between the variants and the originals is also investigated. In speech recognition experiments, the proposed adaptation with the subword variants reduced the word error rate by 12.0% relative in a Japanese conversational broadcast task.

key words: *speech recognition, speech variability, discriminative training, acoustic model*

1. Introduction

The use of automatic speech recognition (ASR) technology has continued to grow in recent years. However, the performance of ASR should be improved for applications such as closed-captioning services of live TV programs [1] and metadata production of archived content [2]. An ASR system that can generate captions in real-time has been used successfully for broadcast news and sports commentaries [1]. However, it has not demonstrated sufficient accuracy under spontaneous or conversational conditions for practical applications.

In this paper, TV programs consisting of spontaneous or conversational speech are our targets for speech recognition. In this kind of speech recognition task, there are phrases and expressions that are specific to a program or a speaking style. Spontaneous speech with such particular wording is often verbose and obscure due to the absence of scripts. Words pronounced indistinctly while figuring out the next words to be spoken often have statistics of acoustic features that are different from training data of an acoustic model. Compared with spontaneous speech, a large amount of read speech is available for training acoustic models for broadcast news. It is, therefore, rational to make use of such acoustic models for recognition tasks involving broadcasts that feature conversations. There is a great need for an acoustic model adaptation method that improves the accuracy of recognizing such task-dependent obscure words.

In the literature, the most common approach used with the pronunciation variants improves recognition accuracy at lexicon level. There, pronunciation variation is usually modeled by adding pronunciation variants to the lexicon [3]–[6]. On the other hand, pronunciation variation can also be represented at the subword level in the acoustic model [4], [7], [8]. Here, a subword is a unit of a hidden Markov model (HMM) composing a word, and it typically corresponds to a phoneme or a triphone in speech recognition. This approach, however, does not obtain a sufficient gain in a large-vocabulary task, because such word-independent subword variants affect all the words with the subwords in a lexicon without regarding the task dependency or its error tendency.

Using word models rather than subword models is a straightforward approach for pronunciation variants at the acoustic model level. The word model is an HMM comprised of more states than a subword HMM. The states of the word model are trained from acoustic features of the word without subword boundaries, so that it can be used to train various speech aspects such as deletion of vowels in rapid utterances. This approach, therefore, also settles the pronunciation issues addressed by the lexical level approach. Typically, word models are only trained for the most frequently occurring words, while subword models can be used for other words in a large vocabulary task [4]. The number of parameters of an acoustic model, however, tends to be very large because word-dependent subwords, i.e. word models, are trained for many function words. This points to the need for an effective means of selecting words involving variants.

In this paper, we focus on task-dependent variants of subwords, which is an approach at the acoustic model level. The method appends HMMs for the variants and directly estimates their HMM statistics from task-dependent adaptation data. HMMs dependent on words involving the subword variants are newly created with this method.

The proposed method selects words that frequently induce recognition errors by using word frequencies and word posteriors from the recognized word lattices for the adaptation data. Then, subwords dependent on the selected words are discriminatively trained by using linear transforms. In the discriminative adaptation, subword accuracy for a mini-

Manuscript received November 30, 2009.
Manuscript revised February 26, 2010.

[†]The authors are with NHK(Japan Broadcasting Corporation) Science & Technology Research Laboratories, Tokyo, 157–8510 Japan.
a) E-mail: satou.s-gu@nhk.or.jp

DOI: 10.1587/transinf.E93.D.2370

mum phoneme error (MPE) criterion is obtained while discriminating between the appended subwords and the original ones. This method is expected to improve recognition accuracy for indistinctly pronounced words or phrases that are dependent on a speaking style specific to the task.

The rest of this paper is organized as follows. Section 2 presents the word identification process of the proposed method, and Sect. 3 describes the discriminative linear transforms we used. Experimental results on a conversational broadcast task are given in Sect. 4 with a discussion. Finally, in Sect. 5 we conclude and outline some directions for the future work.

2. Identification of Speech Variants

A schematic diagram showing the flow of the proposed method is presented in Fig. 1. From an original acoustic model and adaptation data consisting of utterances and transcriptions, the proposed method yields a task-adapted acoustic model by using discriminative adaptation. The adaptation utterances are decoded to produce lattices of word hypotheses used for the proposed identification of words including subword variations. Then the word-dependent subwords are appended to the acoustic model and discriminatively adapted to the task. The following is a detailed procedure of the proposed method.

First, with the proposed identification method, frequent words are selected from the adaptation data given for a task. A set of words $\mathcal{W}^{\mathrm{f}}$,

$$\mathcal{W}^{\mathrm{f}} = \{w \in \mathcal{W}^{\mathrm{r}}; N(w) \geq N^{\mathrm{f}}\}, \tag{1}$$

is selected from transcriptions of the adaptation utterances by using the lower limit of word frequency N^{f}, where $\mathcal{W}^{\mathrm{r}}$ is a set of words observed in the adaptation transcriptions, and $N(w)$ gives word frequency of a word w. Based on only these counts, function words in common with any broadcast task are easily selected, and it is hard to find task-specific words in the most frequent words. Table 1 gives an example of the most frequent words w and their frequencies $N(w)$ observed in a Japanese TV talk show called "Today's Close-up" described in Sect. 4.1. In this paper, we use gender-dependent acoustic models and gender-dependent words by using gender prefixes of "M_" and "F_" which respectively indicate male and female.

Second, words are selected from $\mathcal{W}^{\mathrm{f}}$ if they frequently induce recognition errors by task-dependent subword variants. Expectations of word posteriors of hypothesis lattices are used in this selection. An example of a lattice of word hypotheses and a sequence of reference words are schematically presented in Fig. 2. The word lattice consists of a set of nodes representing points in time and a set of links $\{i, j, a, b, c\}$ representing reference and hypothesis words $\{w_i, w_j, w_a, w_b, w_c\}$. All links are also attached with forward-backward probabilities, i.e. posterior probabilities, $\{p_i, p_j, p_a, p_b, p_c\}$.

These probabilities are calculated by using a forward-probability $\alpha(j)$ and a backward-probability $\beta(j)$. Let $\alpha(\mathcal{B}) = 1.0$ be the forward-probability of the beginning silence link, $\beta(\mathcal{E}) = 1.0$ be the backward-probability of the ending silence link. $L^l(j)$ is the set of preceding links connected to j, $L^r(j)$ is the set of following links connected from j, and $l(j)$ is the likelihood calculated from the acoustic and language scores of link j. p_j, $\alpha(j)$, and $\beta(j)$ are calculated by using a recursive algorithm.

$$\alpha(j) = \sum_{j' \in L^l(j)} \alpha(j') l(j'), \tag{2}$$

$$\beta(j) = \sum_{j' \in L^r(j)} \beta(j') l(j'), \tag{3}$$

Table 1 Example of the 12 most frequent words w and their frequency $N(w)$ observed in a program.

w	$N(w)$	w	$N(w)$	w	$N(w)$
M_の	11,602	M_は	7,243	M_いう	5,692
M_と	10,361	F_の	6,988	M_です	5,032
M_が	8,803	M_に	6,713	M_で	4,401
M_ね	7,266	M_を	6,564	F_と	4,303

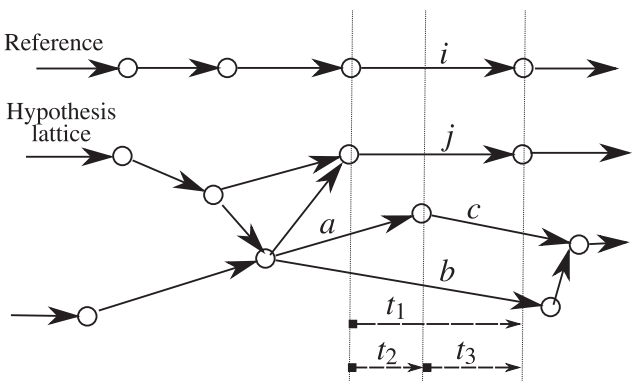

Fig. 2 Example of a lattice of word hypotheses and a sequence of reference words.

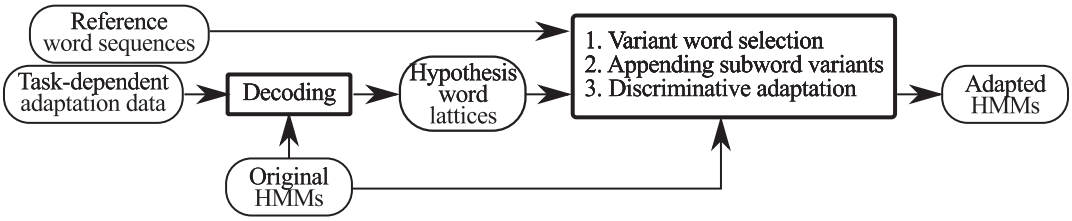

Fig. 1 Schematic diagram of proposed task adaptation.

$$p_j = \frac{\alpha(j)\beta(j)l(j)}{\alpha(\mathcal{E})l(\mathcal{E})}. \quad (4)$$

Let $\mathcal{L}^{\mathrm{R}}$ be a set of links in the reference sequences and $\mathcal{L}^{\mathrm{H}}$ represent a set of links in the hypothesis lattices. Then a posterior probability of a link $j \in \mathcal{L}^{\mathrm{H}}$ at a section occupied by a link $i \in \mathcal{L}^{\mathrm{R}}$ can be calculated as a posterior probability weighted by a ratio of the occupations,

$$\mathcal{P}_i(j) = \frac{T_i(j)}{T(i)} p_j, \quad (5)$$

where $T(i)$ is the number of frames occupied by reference link i, $T_i(j)$ is the number of frames of a reference link i overlapping with a hypothesis link j. For example, the Ts of links shown in Fig. 2 are $T(i) = t_1$, $T_i(j) = t_1$, $T_i(a) = t_2$, $T_i(b) = t_1$, and $T_i(c) = t_3$.

Weighted sum of posterior probabilities of a hypothesis word h in periods occupied by a reference word r are accumulated at $C_r(h)$ over all reference links associated with word r, $\{i \in \mathcal{L}^{\mathrm{R}}; w_i = r\}$, and all hypothesis links associated with h, $\{j \in \mathcal{L}^{\mathrm{H}}; w_j = h\}$, in the training utterances.

$$C_r(h) = \sum_{\{i \in \mathcal{L}^{\mathrm{R}}; w_i = r\}} \sum_{\{j \in \mathcal{L}^{\mathrm{H}}; w_j = h\}} \mathcal{P}_i(j). \quad (6)$$

Note that $\mathcal{P}_i(j)$ is proportional to $T_i(j)$, and $\mathcal{P}_i(j) = 0$ if i and j have no overlap period.

Expectation $\bar{C}_r(h)$ of a word hypothesis h at a period occupied by a reference word r is calculated by marginalizing over a set of words $\mathcal{W}^{\mathrm{H}}$ observed in the hypothesis lattices,

$$\bar{C}_r(h) = \frac{C_r(h)}{\sum_{w \in \mathcal{W}^{\mathrm{H}}} C_r(w)}. \quad (7)$$

Now, $\bar{C}_r(h = r)$ is calculated as an expectation of word r correctly recognized at periods occupied by r. Words with low expectations are regarded as words comprised of subword variations. The proposed method, therefore, selects a set of words $\mathcal{W}^{\mathrm{c}}$,

$$\mathcal{W}^{\mathrm{c}} = \{w \in \mathcal{W}^{\mathrm{f}}; \bar{C}_w(w) \leq C^{\mathrm{s}}\}, \quad (8)$$

where C^{s} is a higher limit of the expectation $\bar{C}_w(w)$ for the word selection.

Assuming that each of the least accurate words $\{w \in \mathcal{W}^{\mathrm{c}}\}$ comprises variants of subwords, word dependent subword models for these words are newly created by copying HMM parameters of their original subword models.

An example of the least accurate words, i.e. words with lower $\bar{C}_w(w)$, are given in Table 2 with their statistics of expectations of posteriors $\bar{C}_w(w)$ and occurrences $N(w)$. These words were automatically selected from the same conversational broadcast TV program as Table 1. The acoustic model used for this selection was trained from read speech of broadcast news. Therefore, compared to a set of words $\mathcal{W}^{\mathrm{f}}$ listed in Table 1, task-specific words that were distinctive to conversational speech were selected by the proposed posteriors $\bar{C}_w(w)$.

Table 2 Example of the 12 least accurate words w and their statistics of $\bar{C}_w(w)$ and $N(w)$.

w	$\bar{C}_w(w)$	$N(w)$	w	$\bar{C}_w(w)$	$N(w)$
M_ま	0.09	1,212	M_あの	0.29	1,914
F_ま	0.14	639	M_いい	0.37	540
F_って	0.19	1,101	M_こう	0.38	785
M_まあ	0.21	603	M_よ	0.38	895
M_えー	0.22	939	M_だ	0.38	631
F_あの	0.24	818	F_いう	0.41	2,651

Table 3 Example of word dependent subwords. For practical use, context-dependent subwords of triphones are generated from these subword sequences.

word	subwords	word	subwords
M_ま	M_m0, M_a0	M_あの	M_a6, M_n6, M_o6
F_ま	F_m1, F_a1	M_いい	M_i:7
F_って	F_Q2, F_t2, F_e2	M_こう	M_k8, M_o:8
M_まあ	M_m3, M_a:3	M_よ	M_y9, M_o9
M_えー	M_e:4	M_だ	M_d10, M_a10
F_あの	F_a5, F_n5, F_o5	F_いう	F_i11, F_u11
			F_y11, F_u:11

Examples of the word-dependent subwords newly created from the least accurate words are given in Table 3. In the table, selected words and given subword sequences are shown in order of their posteriors $\bar{C}_w(w)$. The word-dependent subwords are represented by concatenations of gender prefixes of "M_" and "F_," raw phonemes, and word indices. For a practical use, context-dependent subwords of triphones are generated from the subword sequences.

Since gender-dependent words are used for the selection, gender-dependent subword variants are identified by using the prefixes, "M_" and "F_", as well as task-dependent variants. It is noted that the dependence upon gender can also be easily extended to the dependence upon speaking styles of speakers by extending the prefixes to those indicating speakers or attributes of speakers.

3. Discriminative Adaptation

Discriminative criteria such as MPE [9], [10] have recently been used to improve speech recognition. The subword variants extracted by the proposed method should be appropriately adapted to the acoustic model trained with the discriminative criteria. Adaptation methods based on maximum likelihood (ML) criteria often degrade the recognition performance of discriminative acoustic models because the ML estimations do not consider the relationship between different phonemes. On the other hand, the discriminative models tend to over-fit to the training data, and they may have poor generalization ability on unseen test data. To prevent discriminative models from over-fitting to a rather small amount of adaptation utterances of the task, a transform-based approach can be used.

The proposed method adopts discriminative linear transform (DLT) [11], which is a transform of Gaussian parameters and is estimated based on the discriminative criteria. The same formula as MLLR is used to adapt a Gaussian mean μ_{km} for m-th Gaussian component of HMM state k,

$$\hat{\mu}_{km} = \mathcal{A}\mu_{km} + b = W\xi_{km}, \tag{9}$$

where $\mathcal{A}$ is an $(n \times n)$ mean transform matrix (n is the dimension of the feature), b is a bias vector, $W = [b\mathcal{A}]$, and $\xi_{km} = [1\mu'_{km}]'$ is an extended mean vector.

In the DLT, a transform is estimated in order to maximize the estimate of phoneme accuracy. The objective function of the estimation is

$$\mathcal{F}^{\mathrm{MPE}}(W) = \sum_{s^{\mathrm{h}}} P(s^{\mathrm{h}}|X, W, \Lambda)A(s^{\mathrm{h}}, s^{\mathrm{r}}), \tag{10}$$

where X is an observation sequence of training data, s^{r} is a reference sequence of subwords, s^{h} is a hypothesis sequence of subwords, and Λ is a set of model parameters of an original acoustic model. Here, $A(s^{\mathrm{h}}, s^{\mathrm{r}})$ is the subword accuracy of hypothesis s^{h} and can be efficiently estimated by using subword accuracy $A^{\mathrm{sub}}(q)$ of subword link q composing hypothesis word link $j \in \mathcal{L}^{\mathrm{H}}$.

Assuming that the Gaussian covariances are diagonal, a closed-form solution can be obtained by estimating $\mathbf{w}_n$ of n-th row of matrix W:

$$\mathbf{w}_n = G_n^{-1}k'_n, \tag{11}$$

where the statistics G_n and k_n are given by:

$$G_n = \sum_{km} \frac{1}{\sigma^2_{km(n)}} \left(\gamma^{\mathrm{MPE}}_{km} + D_{km}\right)\xi_{km}\xi'_{km}, \tag{12}$$

$$k_n = \sum_{km} \frac{1}{\sigma^2_{km(n)}} \left(\theta_{km(n)} + D_{km}\tilde{\mu}_{km(n)}\right)\xi'_{km}. \tag{13}$$

Here, $\sigma^2_{km(n)}$ is the n-th element of the diagonal variance, D_{km} is a smoothing factor empirically determined for each Gaussian component. The required statistics for this estimation are defined as below:

$$\gamma^{\mathrm{MPE}}_{km} = \sum_q \sum_t \gamma_{km}(t)\gamma^{\mathrm{MPE}}_q, \tag{14}$$

$$\theta_{km(n)} = \sum_q \sum_t \gamma_{km}(t)\gamma^{\mathrm{MPE}}_q o_n(t), \tag{15}$$

where $o_n(t)$ is the n-th element of the feature at time t and $\gamma_{km}(t)$ is an occupation probability of Gaussian component km at time t.

The key quantity required to optimize the objective function is:

$$\gamma^{\mathrm{MPE}}_q = \frac{1}{\kappa}\frac{\partial \mathcal{F}^{\mathrm{MPE}}}{\partial \log p(q)} \tag{16}$$

$$= \gamma_q(c(q) - c^{\mathrm{avg}}_R), \tag{17}$$

which is the differential of the objective function with regard to the log likelihood $\log p(q)$, for the subword link q, scaled by $\frac{1}{\kappa}$. The quantity is also related to the posterior "occupation probability" γ_q of link q calculated from lattice-based forward-backward algorithm, $c(q)$ is average subword accuracy $A(s^{\mathrm{h}}, s^{\mathrm{r}})$ of subword sequences passing through the link q, and c^{avg}_R is the average subword accuracy of all the subword sequences in the hypothesis lattice for the R-th training utterance.

The phone accuracy $A(s^{\mathrm{h}}, s^{\mathrm{r}})$ required in the calculation of $c(q)$ ideally equals the number of correct phones minus insertions. The exact form of the function, therefore, equivalently expressed as a sum of $\breve{A}^{\mathrm{sub}}(q)$ over all phones q in hypothesis sequence s^h, where $\breve{A}^{\mathrm{sub}}(q)$ is a function giving 1 if q is a correct phone, 0 if q is a substitution, and -1 if q is an insertion. Since the computation of the above expression requires alignment of the reference and hypothesis sequences and this is computationally expensive, $\breve{A}^{\mathrm{sub}}(q)$ is approximated as follows. Given a hypothesis subword q, a subword z in the reference script overlapping in time with q gives the overlapped proportion of the length $e(q, z)$,

$$A^{\mathrm{sub}}(q) = -1 + 2e(q, z), \tag{18}$$

if z and q are the same subword, and otherwise,

$$A^{\mathrm{sub}}(q) = -1 + e(q, z). \tag{19}$$

The value of $c(q)$ is calculated from the subword accuracy $A^{\mathrm{sub}}(q)$ of link q by using an algorithm similar to the forward-backward algorithm [9].

Context-dependent subwords whose center subwords are identical may have too close feature distribution to train discriminatively. Therefore, context-independent subwords, namely monophones, are generally used for the subword accuracy calculation represented by Eq. (18) and (19) even if context-dependent subwords are used for decoding. However, the subword variations extracted and appended by the proposed method may have feature distributions that are very different from the original subwords. The subword accuracy that discriminates between the appended subwords and the original ones may be effective in the discriminative training. Therefore, two subword accuracies, A^{gen} and A^{disc}, are investigated in this paper.

3.1 Subword Accuracy without Discriminating the Variants from the Original

The subword accuracy A^{gen} is calculated in a general way; namely, suffixes of word indices of the subword variations are ignored for the calculation of $A^{\mathrm{sub}}(q)$. For example, Eq. (18) is applied to the calculation of $A^{\mathrm{sub}}(q)$ for a link q associated with the subword "M_m0" at a period occupied by a reference subword "M_m". In this case, $A^{\mathrm{sub}}(q)$ makes γ^{MPE}_q larger than that calculated by Eq. (19). Consequently, the transforms of these subwords are respectively optimized without discounting the occupations of features commonly occupied by these subwords. A^{gen} is, therefore, expected to yield better performance if the proposed subword variants have feature distributions that are very close to the original subwords.

3.2 Subword Accuracy by Discriminating the Variants

In this approach, the subword accuracy A^{disc} is given by dis-

criminating between the appended subwords and the original one. It is calculated from $A^{\mathrm{sub}}(q)$ yielded by using suffixes of word indices. In this case, Eq. (19) is applied to the calculation of $A^{\mathrm{sub}}(q)$ for the link q in the example described above. In contrast to A^{gen}, the transforms are estimated by discounting the occupations of features commonly occupied by these subwords so as to increase the discrimination between these subwords. It is expected that A^{disc} will yield better performance if the subword variants have different feature distributions.

Because the subword variants are dependent on words in the error calculation, recognition errors of subword variants are also regarded as word errors. This method, therefore, partially optimizes parameters of an acoustic model with the minimum word error criteria.

4. Experiments

4.1 Experimental Setup

Experiments were conducted to evaluate the proposed acoustic model adaptation. A conversational broadcast of a Japanese TV talk show called "Today's Close-up," which reports informatively various news stories, was evaluated. The conversation was comprised by reporting of a female newscaster along with various in-studio guests. The evaluation speech data, obtained from seven episodes that aired in May 2008, consisted of 12,356 words uttered by ten speakers. Two of the speakers "overlapped" between the evaluation set and the adaptation utterances described below. One was the female newscaster, who uttered 3,369 words of the evaluation set. The other was a male reporter, who uttered 385 words.

The n-gram language models used in this experiment were word bigrams for the first pass in the continuous speech recognition and trigrams for the second pass. By using linear interpolation, language models trained from broadcast news (143 M words), informative reporting (38 M words), and transcriptions of press conferences (44 M words) were combined into the model used for the evaluation. There were 100 K vocabulary words, and the trigram language model showed a perplexity of 87, with an out-of-vocabulary rate of 0.3% against the evaluation data.

Gender-dependent acoustic models as original models to be adapted were trained from NHK's Japanese broadcast news data consisting of 340 hours of male utterances and 250 hours of female utterances. Each gender-dependent model consisted of about 4 K clustered states with 16 Gaussian mixtures for triphone HMMs with three emission states. Parameters of the models were estimated using MPE criteria gender-dependently. The gender dependent models were merged by using gender prefixes of the subwords for the adaptation. From 248 episodes of the task program, 31 hours of utterances were transcribed for the adaptation. The transcription consisted of word and gender labels. With these labels and gender prefixes "M_" and "F_" for subwords, the gender-dependent acoustic models were merged into one set of models and were adapted all at once.

The conventional acoustic model was adaptively trained without adding any HMMs of subword variants. The shared 7,941 states of triphone HMMs were clustered according to their gender-dependent center phonemes. The linear transforms of the state parameters were estimated for these regression clusters as well as a shared silence. Specifically, there were 81 regression clusters consisting of 40 male phonemes, 40 female phonemes, and a silence.

For the variant words selected by the proposed method, the triphone HMMs, and the states referred to by the triphones were created by copying the parameters of their originals. Then, their word indices were given to the labels of the HMMs and the states as shown in Table 3. Here, state sharing was kept if some states were shared by the triphone HMMs composing the same word. The regression clusters of the variants were also created by clustering the states according to their center phonemes, which include word indices, so that the states referred to by triphones with a common center phoneme were transformed by the same matrix. The number of additional HMM states was, therefore, not proportional to the number of additional regression clusters.

The smoothing value D_{km} for the DLT estimation was chosen as follows,

$$D_{km} = 2.5 \sum_{q} \sum_{t} \gamma_{km}(t) \max(0, -\gamma_q^{\mathrm{MPE}}). \tag{20}$$

A continuous speech recognizer [12] with parallel gender-dependent acoustic models was used for the evaluation. The decoder searches a phonetic lexicon tree for each gender in parallel, but with a common beam threshold. Search transitions between male and female are allowed in frames when the gender attribute changes in the preceding speech detection based on a dual-gender phoneme recognizer.

4.2 Comparison of Adaptation Methods

With the proposed method, the selection parameter of the lower limit of a word frequency $N^{\mathrm{f}} = 1,000$ extracted 50 words of $\mathcal{W}^{\mathrm{f}}$ from the adaptation utterances. Numbers of extracted words (#words), numbers of regression clusters (#reg), and numbers of HMM states (#states) of the proposed method are given in Table 4. The proposed acoustic model generated by the higher limit of the posterior $C^{\mathrm{s}} = 0.7$, $C^{\mathrm{s}} = 0.8$, and $C^{\mathrm{s}} = 0.9$ are compared with the conventional acoustic model. By increasing C^{s}, numbers of the acoustic model parameters were increased. When C^{s}

Table 4 Increase in number of parameters due to the proposed method ($N^{\mathrm{f}} = 1,000$).

method	#words	#reg	#states
conventional	0	81	7941
proposed $C^{\mathrm{s}} = 0.7$	31	153	8166
proposed $C^{\mathrm{s}} = 0.8$	45	200	8337
proposed $C^{\mathrm{s}} = 0.9$	50	221	8417
Only $\mathcal{W}^{\mathrm{f}}$ ($C^{\mathrm{s}} = 1.0$)			

was 0.9, all 50 words of $\mathcal{W}^{\mathrm{f}}$ were extracted. The proposed method with $C^{\mathrm{s}} = 0.9$ and $C^{\mathrm{s}} = 1.0$, therefore, created new 476 HMM states.

Figure 3 compares the word error rates (WERs) of the conventional method, the proposed method based on A^{gen}, and the proposed method based on A^{disc}. This figure shows the result of $C^{\mathrm{s}} = 0.7$ and the WERs are plotted with iterations of parameter updates, since DLTs are estimated in an iterative manner. A WER plotted at 0 in the iteration is a result yielded by the acoustic model without adaptation. The estimation was repeated six times, beyond which no improvements were observed.

The proposed method reduced WERs compared with the conventional method for both subword accuracy A^{gen} and A^{disc}. The largest improvement was yielded by the proposed method based on A^{disc}. In this case, WER was 19.6% with a word error reduction rate of 10.9% compared to the WER of 22.0% by the model without adaptation.

As described in Sect. 3, there are two possible reasons why A^{disc} showed a slightly lower WER than A^{gen}. One is the overlap of the feature distributions observed between the subword variants and the original subwords. Improvement of discriminative ability between them might made the recognition accuracy better. The other reason is a criteria used for the estimation of the HMM parameters. As described in Sect. 3.2, the proposed method based on A^{disc} partially trains transforms of words likely to be falsely recognized with minimum word error criteria. Such a criteria might reduce the word errors.

Recognition results of the proposed method extracting word variants with $C^{\mathrm{s}} = 0.8$ and $C^{\mathrm{s}} = 0.9$ are respectively given in Fig. 4 and Fig. 5 in a similar manner with Fig. 3. As C^{s} was increased from 0.7 to 0.9, more confident words, which were correctly recognized, were selected for a set of word variants $\mathcal{W}^{\mathrm{c}}$. By increasing the number of these confident words in $\mathcal{W}^{\mathrm{c}}$, the difference in WER between A^{gen} and A^{disc} became smaller. It is expected that the subword variations appended by these confident words have feature distributions that were very close to the original subwords while the subword variants of less confident words had different distributions from the original ones. If the calculation of subword accuracy had been optimized according to a word confidence $\bar{C}_w(w)$, i.e. A^{disc} had been applied to falsely recognized words and A^{gen} had been applied to confident words, more improvement might be yielded. Further investigation is necessary to clarify the exact reason of the difference between A^{gen} and A^{disc}.

The largest improvement in these experiments was yielded when $C^{\mathrm{s}} = 0.8$, and WER was 19.3% with a word error reduction rate of 12.3%. The $\mathcal{W}^{\mathrm{c}}$ for $C^{\mathrm{s}} = 0.9$ was identical with the set of words $\mathcal{W}^{\mathrm{f}}$ consisting of the 50 words selected by only the lower limit of word frequency $N^{\mathrm{f}} = 1,000$. Subword variants created for the set of words $\mathcal{W}^{\mathrm{f}}$ ($C^{\mathrm{s}} = 1.0$ and $C^{\mathrm{s}} = 0.9$) showed WER of 19.6% and did not reduce WER compared to the proposed word set $\mathcal{W}^{\mathrm{c}}$ selected by $C^{\mathrm{s}} = 0.7$ or $C^{\mathrm{s}} = 0.8$. Decreasing the higher limit of the posterior $C^{\mathrm{s}} = 0.7$, the proposed method created only

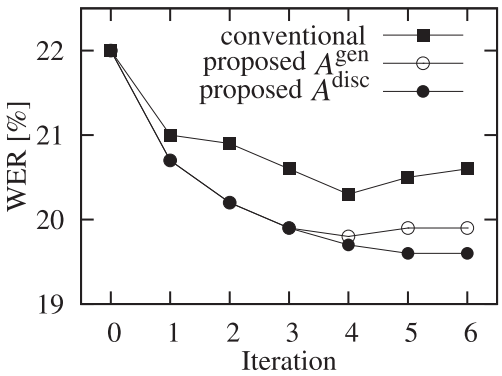

Fig. 3 Comparison of WER with different adaptation methods ($N^{\mathrm{f}} = 1,000$, $C^{\mathrm{s}} = 0.7$).

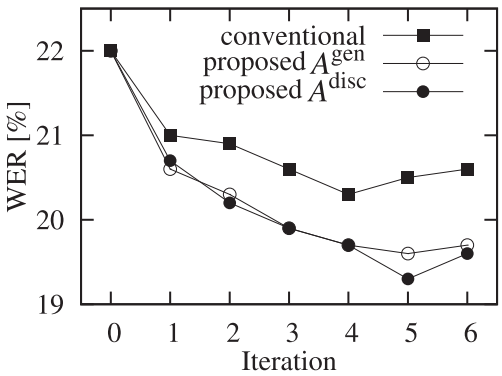

Fig. 4 Comparison of WER with different adaptation methods ($N^{\mathrm{f}} = 1,000$, $C^{\mathrm{s}} = 0.8$).

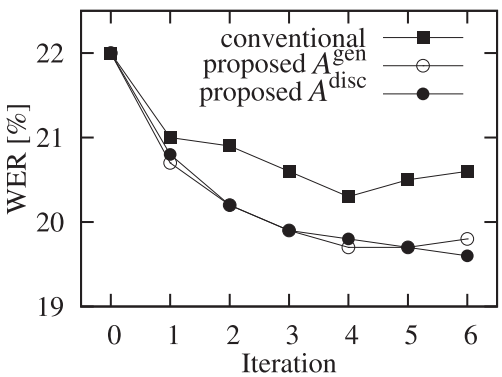

Fig. 5 Comparison of WER with different adaptation methods ($N^{\mathrm{f}} = 1,000$, $C^{\mathrm{s}} = 0.9$ and $C^{\mathrm{s}} = 1.0$).

225 states and 47% of HMM states (476 states) by only the lower limit of word frequency $N^{\mathrm{f}} = 1,000$, and it efficiently selected the subword variants to achieve the improvement. It is also noted that $C^{\mathrm{s}} = 0.8$ reduced the WER by creating 83% of additional states of $\mathcal{W}^{\mathrm{f}}$. It is confirmed that the proposed word selection is effective by introducing the word confidence $\bar{C}_w(w)$ in the discriminative acoustic model adaptation.

In the following section, A^{disc} was used for the adaptation experiments and iterative parameter estimation was

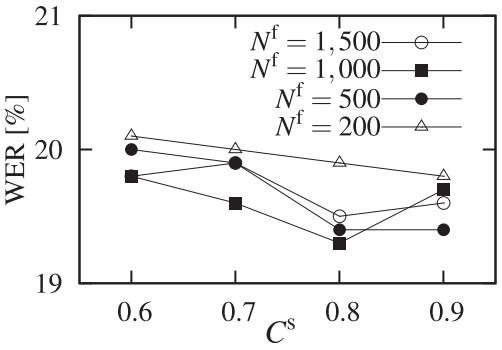

Fig. 6 Comparison of WER with different N^f.

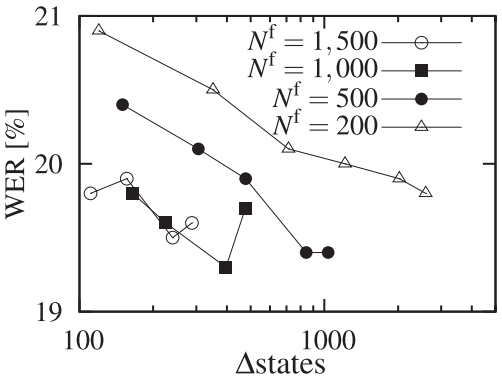

Fig. 7 WERs yielded by numbers of HMM states created for subword variants.

repeated 5 times.

4.3 Comparison of Increased Number of Model Parameters

Figure 6 compares WERs yielded by four different N^fs of 1,500, 1,000, 500, and 200. WERs are plotted against the value of C^s from 0.6 to 0.9. By decreasing the number of the lower limit of a word frequency N^f from 1,000 to 200, WERs with $C^s = 0.8$ were increased because of over training of newly created HMM states. Since occupation counts were allocated between additional states and the original states, accuracies of estimated parameters of the original states might be degraded by an excessive supplement of subword variants. Subword variants with $N^f = 1,500$ and $C^s = 0.8$ also increased WERs because smaller numbers of subword variants were extracted than those of optimal variant extraction. WERs with $C^s = 0.9$ except for $N^f = 200$ were not improved compared with those with $C^s = 0.8$. The proposed method, therefore, achieved lower WERs for $C^s = 0.8$ with smaller numbers of additional HMM states than those for $C^s = 0.9$.

Changing the value of C^s, Fig. 7 plots WERs against Δ states, which are numbers of HMM states newly created for subword variants. When Δ states are less than 400, $N^f = 1,000$ achieved almost the lowest WERs in those achieved by the other N^fs at the same Δ states. It is confirmed that the proposed method of $N^f = 1,000$ and $C^s = 0.8$ efficiently created subword variants. This result had a statistically significant level of 0.03 in a matched pairs test [13] in comparison with the result of the word variants selected without the parameter C^s, which created 20% more variant states.

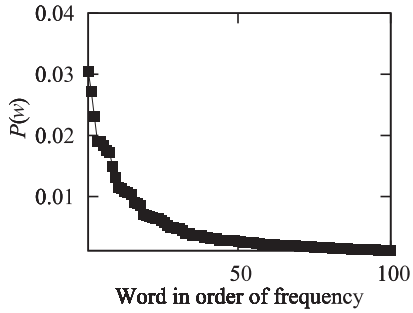

Fig. 8 Word probabilities $P(w)$ of the 100 most frequent words in the adaptation data.

Figure 8 shows the word probabilities $P(w)$ of the 100 most frequent words in the adaptation data. In this experiment, $N^f = 1,000$ was equivalent to $P(w) > 0.003$ and $P(\mathcal{W}^f)$ was 0.42. $P(\mathcal{W}^f)$ selected by $N^f = 500$ was 0.5. Figure 7 shows that the additional HMM states of words selected by $N^f = 500$ did not reduce the WER because the $P(w)$s of the words were rather small. If the distribution of $P(w)$s is close to that of Fig. 8, it is an indication that the proposed method improves the recognition as a result of setting N^f so that $P(\mathcal{W}^f)$ is about 0.4.

On the other hand, such an issue of over-training as a result of $N^f = 200$ is dependent on the amount of adaptation data and number of parameters composing the acoustic model. Finding the optimal value of N^f to prevent over-training requires an effective method for estimating the adequacy of models' discriminative ability for given adaptation data.

In this experiment, A^{disc} in the proposed method was applied to both subwords, to which A^{gen} or A^{disc} should have been applied. The ratio of these sets of subwords is dependent on the speaking style of the given task. The optimal value of $C^s = 0.8$, therefore, may change with the ratio. However, C^s may be easily optimized if the subword accuracy A^{sub} is adequately selected according to the discriminative ability of $\bar{C}_w(w)$, as mentioned in Sect. 4.2. The selection of A^{sub} would also improve recognition accuracy.

4.4 Improvements of Variant Words

The error rates of each adaptation method are listed in Table 5 in order to compare the improvements regarding the 45 variant words selected with $C^s = 0.8$ and the other words. WERs for all the evaluation data are compared with keyword error rates (KER) under the assumption that the 45 variant words are our objective keywords. The error reductions relative to the acoustic model without adaptation are also shown. Conventional adaptation, which has no addi-

Table 5 Error rates and error reductions for all evaluation words and variant words. [%]

Adaptation	WER	Reduction	KER	Reduction
w/o adaptation	22.0	0	28.2	0
conventional	20.5	6.8	26.4	6.4
proposed	19.3	12.3	24.4	13.5

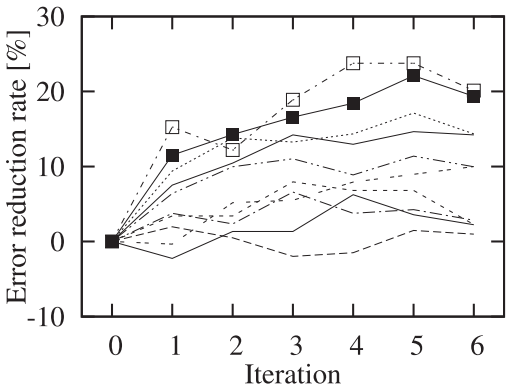

Fig. 9 Error reductions of speakers relative to the original acoustic model.

tional subword variants, achieved a 6.4% keyword error reduction. On the other hand, the proposed method achieved a 13.5% keyword error reduction.

The conventional method achieved 6.8% of word error reduction for all the evaluation data and the proposed method reduced 12.3% of error words. This reduction was close to one calculated by only the keywords. Besides features occupied by corresponding subwords of references, features occupied by subwords of error hypothesis are used for the discriminative estimation of HMM parameters. It is, therefore, considered that error words besides errors occurring on the selected 45 words were reduced by the discriminative training of the proposed subword variants.

4.5 Comparison of Improvements of Speakers

As described in Sect. 4.1, two closed evaluated speakers were also included in the adaptation data. To confirm the improvement of the open speakers, the error reduction rates for each speaker are plotted in Fig. 9. This figure shows the reduction rates relative to the original acoustic model. The result of the closed speakers are plotted with squares, where the black ones show the results of the female newscaster, and the white ones show the result of the male reporter. It was confirmed that WERs of all the open speakers were reduced by the proposed acoustic model adaptation by repeating the discriminative estimation more than four times. The maximum error reduction rate of 18% was yielded from one of the open speakers and minimum error reduction rate of 1.4% was yielded from another open speaker.

Improvements relative to the conventional adaptation are also confirmed in Fig. 10, which shows the error reductions for each speaker. Speaker IDs of 1 and 2 are the result of the closed speakers: the news caster and the reporter,

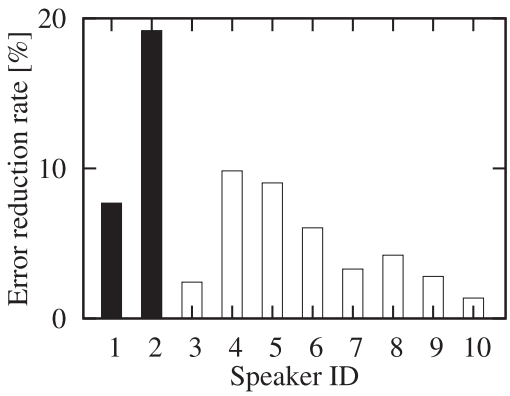

Fig. 10 Error reductions of speakers relative to the conventional adaptation.

respectively. The maximum error reduction rate of 9.8% was yielded from one of the open speakers and minimum error reduction rate of 1.3% was yielded from another open speaker.

It is confirmed that the proposed method extracted not only speaker-dependent subword variants but also speaker-independent subword variants.

5. Conclusion

This paper proposed a new discriminative method of acoustic model adaptation that deals with a task-dependent speaking style and acoustic variants. To improve recognition accuracy against indistinctly pronounced words, the method newly creates subwords dependent on words inducing recognition errors by using recognized word lattices for the adaptation data.

A transcription experiment implementing this method for a Japanese conversational broadcast showed that the proposed adaptation reduced error words by 12.3% relative to the original acoustic model and 5.9% relative to the conventional adaptation. It is also confirmed that the proposed method efficiently achieved lower WERs with smaller numbers of additional HMM states than those achieved by subword variants with only the lower limit of the word frequency.

Future efforts will involve optimizing discriminative abilities of subword variants and extending dependencies of subword variants. As described in Sect. 4.2, the proposed method is expected to yield more improvement if the method appropriately choose A^{gen} or A^{disc} as $A^{\mathrm{sub}}(q)$ dependently on the corresponding words. The most feasible method could be to calculate the accuracy $A^{\mathrm{sub}}(q)$ according to word confidence $\bar{C}_w(w)$. While this paper identifies subword variants observed in words independently of their context, parameters of HMMs for subword variants are expected to be estimated more accurately by identifying subword variants dependent on their word context in commonly observed phrases.

References

[1] T. Imai, A. Kobayashi, S. Sato, K. Onoe, and T. Kobayakawa, "Speech recognition for subtitling Japanese live broadcasts," 18th International Congress on Acoustics, pp.165–168, 2004.
[2] M. Sano, Y. Kawai, H. Sumiyoshi, and N. Yagi, "Metadata production framework and metadata editor," MULTIMEDIA '06: Proc. 14th Annual ACM International Conference on Multimedia, pp.789–790, 2006.
[3] T. Imai, A. Ando, and E. Miyasaka, "A new method for automatic generation of speaker-dependent phonological rules," Proc. ICASSP, pp.864–867, 1995.
[4] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: a survey of the literature," Speech Commun., vol.29, no.24, pp.225–246, 1999.
[5] H. Strik, "Pronunciation adaptation at the lexical level," Proc. ISCA ITRW Workshop Adaptation Methods for Speech Recognition, pp.123–130, 2001.
[6] M. Wester, "Pronunciation modeling for ASR – Knowledge-based and data-derived methods," Comput. Speech Lang., vol.17, pp.69–85, 2003.
[7] R. Singh, B. Raj, and R.M. Stern, "Automatic generation of subword unit for speech recognition systems," IEEE Trans. Speech Audio Process, vol.10, no.2, pp.89–99, 2002.
[8] A. Hämäläinen, L. ten Bosch, and L. Boves, "Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider," Speech Commun., vol.51, no.2, pp.130–150, 2009.
[9] D. Povey, Discriminative Training for Large Vocabulary Speech Recognition, Ph.D. Thesis, Cambridge University Engineering Dept., 2003.
[10] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition," IEEE Signal Process. Mag., vol.25, no.5, pp.14–36, 2008.
[11] L. Wang and P.C. Woodland, "MPE-based discriminative linear transforms for speaker adaptation," Comput. Speech Lang., vol.22, no.3, pp.256–272, 2008.
[12] T. Imai, S. Sato, S. Homma, K. Onoe, and A. Kobayashi, "Online speech detection and dual-gender speech recognition for captioning broadcast news," IEICE Trans. Inf. & Syst., vol.E90-D, no.8, pp.1286–1291, Aug. 2007.
[13] L. Gilick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," Proc. ICASSP, pp.532–535, 1988.

Shoei Sato received a B.E. degree and a M.E. degree in 1993 from Tohoku University. He also received the Dr. Eng. degree from Waseda University in 2008. Since 1995 he has been with NHK Science and Technology Research Laboratories and engaged in the research on digital satellite broadcasting system. He is currently engaged in automatic speech recognition research.

Takahiro Oku received a B.E. degree and a M.E. degree in 2003 from Keio University. He joined Japan Broadcasting Corporation (NHK) in 2003. Since 2007, he has been with the Science and Technology Research Laboratories of Japan Broadcasting Corporation, where he engaged in speech recognition research.

Shinichi Homma received his B.E. degree in electronic and communication engineering from Waseda University, Tokyo, Japan, in 1992. He joined Japan Broadcasting Corporation (NHK) in 1992. Since 1998 he has been with the Science and Technology Research Laboratories, where he has engaged in the research on speech recognition.

Akio Kobayashi received his B.E. degree in 1991 from Waseda University. He joined NHK the same year and has been with the Science and Technical Research Laboratories from 1996. He is currently engaged in speech recognition research. He is a member of the Acoustic Society of Japan (ASJ), the Information Processing Society of Japan, and the Association of Natural Language Processing of Japan.

Toru Imai received his B.E. degree in electrical engineering in 1987 and his Ph.D. degree in information and science in 1999 from Waseda University. He joined Japan Broadcasting Corporation (NHK) in 1987. Since 1990, he has been with the Science and Technology Research Laboratories of NHK. In 1996 he was a visiting scientist with BBN in Massachusetts. He is currently a senior research engineer of the Human & Information Science Research Division at the Laboratories. He is engaged in speech recognition research for broadcasting. He is a member of the Acoustical Society of Japan, the Information Processing Society of Japan, the Institute of Image Information and Television Engineers, and the IEEE.