Speaker Adaptation Intensively Weighted on Mis-Recognized Speech Segments

Takahiro Oku, Yuya Fujita, Akio Kobayashi, and Toru Imai
NHK (Nippon Hoso Kyokai; Japan Broadcasting Corp.)
Science and Technology Research Laboratories, Setagaya 157-8510, Tokyo, Japan
E-mail: {oku.t-le, fujita.y-gc, kobayashi.a-fs, imai.t-mq}@nhk.or.jp Tel: +81-3-5494-3369

***Abstract*— A "re-speak method" is an effective speech recognition method for simultaneous closed-captioning of live broadcasting programs picked up in noisy environments featuring spontaneous or emotional commentary. An acoustic model of the re-speaker needs to be constantly adapted according to the re-speaker's daily health condition or level of fatigue. In this paper, we propose efficient speaker adaptation for the re-speak method. Conventional speaker adaptation is performed uniformly over entire speech segments. In comparison, our proposed speaker adaptation determines intensive adaptation segments corresponding to recognition error parts by comparing speech recognition results and manually error-corrected results. These results are provided in real time by the simultaneous closed-captioning process. Then, the frame-level statistics for speaker adaptation are multiplied by larger weights in proportion to the degree of the recognition errors more over the intensive adaptation segments than they are over the other segments. In an experiment on an information variety program in Japanese broadcasting, our speaker adaptation method reduced the word error rate relatively by 3.4% compared with the conventional uniform adaptation method.**

I. INTRODUCTION

Real-time closed-captioning of live broadcast programs is of great value to the hearing-impaired and elderly people. Although Japanese stenographic keyboards can be used for the real-time captioning, they require six highly skilled operators working at the same time to deal with the great number of homonyms present in ideograms (Kanji). To enable speech to be transcribed more efficiently, we have done extensive research on automatic speech recognition (ASR) for providing closed-captioned TV programs in real time.

We have developed a hybrid ASR system that combines the "direct method," whose input is the original program sound, and the "re-speak method," where another speaker listening to the original speech of a program rephrases the commentary so that it can be recognized for captioning [1][2]. In addition, the re-speak method is especially effective for the closed-captioning of programs such as variety shows because this method can deal with noisy environments and spontaneous or emotional commentary. In this paper, we focus on information variety programs in Japanese broadcasting that are entirely captioned by means of the re-speak method.

In the re-speak method, an acoustic model of the re-speaker needs to be constantly adapted according to the re-speaker's daily health condition or level of fatigue. Therefore, there is a compelling need for efficient speaker adaptation not only beforehand but also in real time.

In the simultaneous captioning process, recognition errors of the re-speaker's speech are immediately detected and manually corrected by an operator with a keyboard. Then, these corrected results are finally broadcasted as closed-captioning. Therefore, speech recognition results that contain recognition errors and manually corrected results are available in real time.

Speaker adaptation of the acoustic model uses these recognition results or manually corrected results and the corresponding re-speaker's speech. Conventional adaptation is performed uniformly over whole re-speaker's speech segments. In comparison, our proposed method is aimed to intensively adapt the acoustic model, especially over the error-corrected parts, to try and not make the same recognition errors again. First, we decide intensive adaptation segments by comparing the recognition results and the manually corrected results with forced alignment on the speech. Then, the acoustic model is adapted more intensively over these segments than it is over the other segments.

Re-speaking is usually performed by two or three speakers alternating about every 20 minutes in consideration of the fatigue caused by continuous re-speaking. Our proposed method adapts the re-speaker's acoustic model at every speaker change by using maximum likelihood linear regression (MLLR) [3] and maximum a posteriori (MAP) [4]. The frame-level statistics for model parameter estimation in MLLR and MAP are multiplied by larger weights over the intensive adaptation segments. Similar frame-weighted HMM training methods were proposed in [5][6]. However, these methods are based on boosting techniques that need iterative processing to decide the weights and are not suitable for our situation of speaker adaptation for simultaneous closed-captioning. On the contrary, our proposed method can immediately and simply decide the weights by comparing speech recognition results and manually corrected results.

The rest of this paper is organized as follows. In Section II, our overall procedures for speaker adaptation by using re-speaker's utterances are described. In Section III, our intensive speaker adaptation method by using manually error-corrected speech recognition results is presented. In Section IV, the experimental setup for performing our speaker

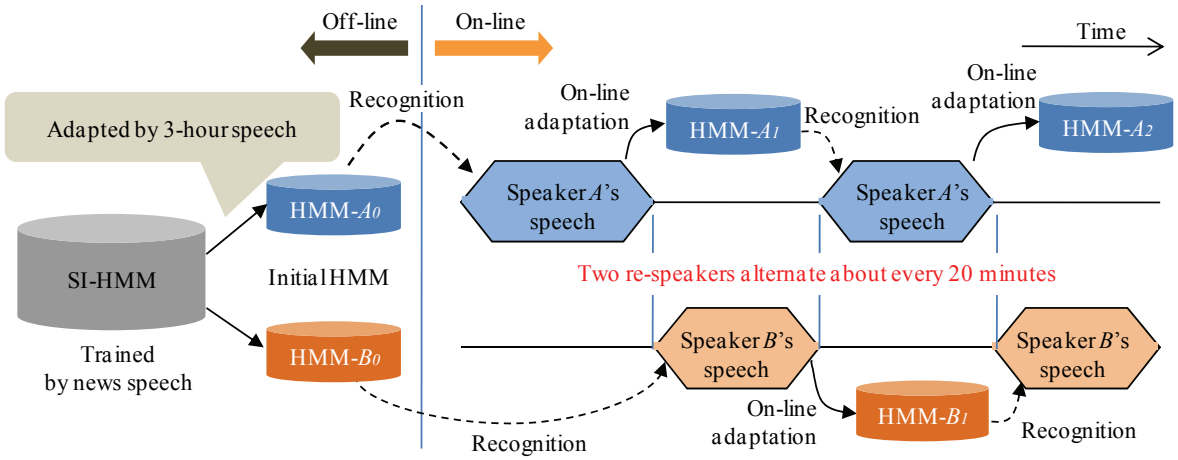

Fig.1 Proposed adaptation method

adaptation on an information variety program in Japanese broadcasting and the experimental results are presented. Section V concludes the study, and future works are mentioned.

II. Speaker Adaptation by Using Re-speaker's Utterances

Our speaker adaptation method is shown schematically in Fig. 1. This figure shows a case of two re-speakers (speaker *A* and speaker *B*). They are female in this paper, and the re-speaking is usually performed for about 20 minutes, at which point, the re-speakers alternate in consideration of the fatigue caused by continuous re-speaking. An original speaker independent acoustic model (SI-HMM) is trained from NHK's Japanese broadcast news. The SI-HMM is adapted to two initial acoustic models (initial HMMs) for each speaker by using past three-hour off-line speech of each re-speaker before the broadcast. It is called off-line adaptation in this paper. At the beginning of the broadcast, the re-speaker's speech is recognized by the initial HMMs. These acoustic models are adapted at every speaker change by using the re-speaker's on-line speech, the recognition results, and the manually error-corrected results before the changes. We call it on-line adaptation. The adapted acoustic model is then used for the recognition when the corresponding re-speaker speaks the next time. Recognized fillers of re-speaking are automatically removed before the error-correction for the broadcast. These fillers are, however, used as training labels for the adaptation of the acoustic models in our experiment.

III. Intensive Speaker Adaptation

In the simultaneous closed-captioning process for live TV programs, speech recognition results containing errors and manually error-corrected recognition results are provided in real-time. Supervised speaker adaptation of the acoustic model is performed by using these error-corrected results. Our proposed method is aimed to intensively adapt the acoustic model, especially over the error-corrected parts, to try and not make the same recognition errors again.

A. Alignment

Our proposed method performs forced alignment on the re-speaker's speech by using recognition results and error-corrected results. The acoustic log likelihoods of the most likely HMM state sequences at frame t generated from the forced alignment by using the recognition results (hypotheses) and the error-corrected results (references) are expressed as $L^H(t)$ and $L^R(t)$, respectively, in this paper.

B. Deciding intensive speaker adaptation segments

The speech segments where the difference between acoustic log likelihood $L^H(t)$ and $L^R(t)$,

$$\Delta L(t) = L^H(t) - L^R(t), \tag{1}$$

is positive are decided to be intensive speaker adaptation segments because frame t is likely to belong to error-corrected parts in these speech segments. In addition, this $\Delta L(t)$ is considered to be the degree of recognition error. If the acoustic model is more intensively adapted in proportion to $\Delta L(t)$ over these segments than it is over the other segments, the accuracy of speech recognition is expected to be improved. In our proposed method, we use recognition results to decide the intensive speaker adaptation segments not only by using the trigram language model but also by using the bigram language model in order to intentionally generate more recognition errors and reasonably expand the intensive adaptation segments.

C. Intensive adaptation

The re-speaker's acoustic model is adapted by using MLLR and then MAP [7]. The posterior probability $\gamma_t(i,m)$, where t, i, and m represent frame, HMM state, and Gaussian mixture component, respectively, calculated in the Baum-Welch algorithm [8] is multiplied by a larger weight in proportion to $\Delta L(t)$ over the intensive adaptation segments by using following equation (2), as shown in Fig. 2.

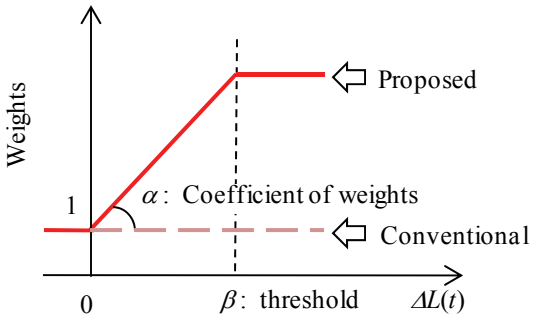

Fig. 2 Weights of intensive speaker adaptation

$$\gamma_t'(i,m)=\begin{cases}\gamma_t(i,m) & (\Delta L(t)\le 0)\\ (\alpha\cdot\Delta L(t)+1.0)\gamma_t(i,m) & (0<\Delta L(t)\le\beta),\\ (\alpha\cdot\beta+1.0)\gamma_t(i,m) & (\beta<\Delta L(t))\end{cases} \quad (2)$$

where α is a coefficient for the weight of intensive speaker adaptation and β is a threshold. Over the non-intensive adaptation segments where $\Delta L(t)$ is not positive, conventional uniform adaptation is performed. This way, the weight is immediately and simply decided by using $\Delta L(t)$ calculated from forced alignment outcomes, and it is suitable for our situation of speaker adaptation for simultaneous closed-captioning.

There is difference in terms of the recording days between past three-hour-long off-line speech for adaptation of the SI-HMM and the re-speaker's on-line speech. Therefore, in our proposed method, the coefficient α for on-line adaptation at every speaker change was larger than the one for off-line adaptation of the SI-HMM.

IV. Experiment

A. Experimental setup

Our proposed speaker adaptation method was evaluated by using two female re-speaker's speech for a Japanese TV information program called "Morning Market," which features conversation between three hosts and various guests in studio. This evaluation set, obtained from six episodes that aired in September and October of 2010, comprised 57,674 words.

The SI-HMM as an original acoustic model to be adapted was trained from NHK's Japanese broadcast news, which consisted of 250 hours of female utterances. The acoustic model consisted of about 4K clustered states with 16 fixed Gaussian mixtures for triphone HMMs with three emission states and three self-transition states.

By using linear interpolation, language models trained from transcription of the TV program (412M words) and the text related to each episode (an average of 17K words) were combined into the model used for the evaluation.

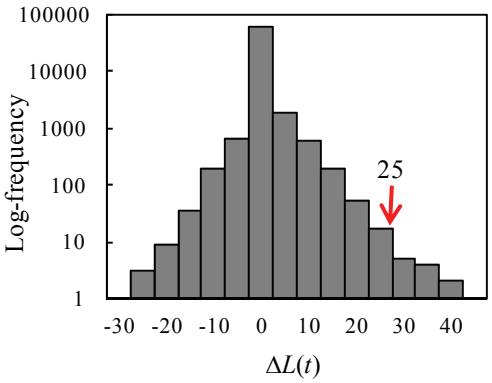

Fig. 3 Distribution of $\Delta L(t)$

TABLE I
WEIGHT α FOR THE INTENSIVE SPEAKER ADAPTATION

	Trigram		Bigram	
	Off-line adaptation	On-line adaptation	Off-line adaptation	On-line adaptation
MLLR	20.0	100.0	10.0	20.0
MAP	0.3	1.5	0.3	0.6

The recognition engine had a 2-pass decoder that uses bigrams and trigrams in the first and second passes, respectively [9].

In our experiment, we compared the following three adaptation methods.

1. Baseline: Speaker adaptation by using normal MLLR and then MAP
2. Trigram: Using recognition results of the decoder's second pass to decide the intensive speaker adaptation segments (proposed method)
3. Bigram: Using recognition results of the decoder's first pass to decide the intensive speaker adaptation segments (proposed method)

Here, we used two regression classes (silence and speech) for MLLR adaptation.

We compared the forced alignment outcomes with recognition results and those with error-corrected results in terms of the past three-hour-long speech of each re-speaker. As a result, the rate at which equation (1) was positive was 5.3% and 4.8% for each female re-speaker in the trigram case and 6.1% and 5.3% in the bigram case. We decided that the threshold β was 25.0 on the basis of the distribution of $\Delta L(t)$ shown in Fig. 3. The rate at which $\Delta L(t)$ was smaller than 25.0 was over 99%. By using re-speaker's speech of the information TV program "Morning Market" that aired in July 2010 as a development set, we decided the coefficient α for the intensive speaker adaptation. The coefficient α in each adaptation method is shown in Table I. Here, this development set is composed of two re-speakers, the same as the evaluation set.

TABLE II
OFF-LINE ADAPTATION WER(%)

	Baseline	Intensive adaptation (Proposed)	
		Trigram	Bigram
Speaker A	9.1	9.1	9.2
Speaker B	10.8	10.6	10.6
All	9.9	9.8	9.8

TABLE III
ON-LINE ADAPTATION WER(%)

	Baseline	Intensive adaptation (Proposed)	
		Trigram	Bigram
Speaker A	8.0	7.9	7.8
Speaker B	10.0	9.6	9.6
All	**8.9**	8.7	**8.6**

B. Results

Tables II and III show comparisons of the word error rates (WERs) for normal speaker adaptation and our proposed intensive speaker adaptation. Table II shows the case for off-line only by using the past three-hour-long speech of each re-speaker. In this case, adaptation at every speaker change was not performed, and the initial HMMs were used for the speech recognition over the whole evaluation set. Table III shows the case for on-line adaptation. In this case, adaptation at every speaker change was performed in addition to off-line adaptation. As a result, in the off-line only case, the WER was slightly improved from 9.9% of the baseline adaptation to 9.8% of both trigram and bigram adaptation. In the on-line adaptation case, the WERs of the baseline and bigram adaptation were 8.9% and 8.6%, respectively (a word error reduction rate of 3.4%). These results confirm that our proposed intensive speaker adaptation method is more effective in the case of on-line adaptation at every speaker change than in the case of off-line adaptation. Furthermore, it was confirmed that compared with trigram, bigram, where the linguistic model constraint is eased, is more effective for improving the WER. We also performed an experiment on a unigram case where recognition results by using the unigram language model were used to decide the intensive speaker adaptation segments, but the WER was worse than the baseline. This is because too many recognition errors were generated by the unigram language model than in the trigram and bigram cases, and the intensive speaker adaptation segments were excessively expanded.

In a discriminative learning based on the minimum phone error (MPE) criterion, statistics on a phoneme lattice are weighted phoneme-arc wisely so as to find system-wise optimal by using large quantities of data. Our proposed method, on the other hand, estimates model parameters to avoid repetitive recognition errors occurred in a similar context of the topic. Our method, thus, weights frame-wisely so as to restore likelihoods locally degraded in the desired hypothesis by using small data.

V. CONCLUSION

A supervised speaker adaptation intensively weighted on mis-recognized speech segments was proposed. Intensive adaptation segments were determined by comparing the recognition results and the manually error-corrected results with forced alignment on re-speaker's speech. To get the recognition results, the bigram language model was used in order to intentionally generate more recognition errors and expand the intensive adaptation segments. Frame-level statistics for MLLR and MAP are multiplied by larger weights in proportion to the degree of the recognition errors over the intensive adaptation segments. Experiments were performed on re-speaker's speech of a TV information program. Our proposed adaptation method reduced the word error rate relatively by 3.4% on the basis of recognition results by using the bigram language model. It was confirmed that our method is more effective in the case of on-line adaptation rather than off-line adaptation. In this paper, the adaptation was performed at every speaker change. Future work will involve further improving the speech recognition accuracy by adaptation not at every speaker change but every several utterances and seeking a method to more reasonably expand the intensive adaptation segments not only by bigram language model.

REFERENCES

[1] S. Homma *et al*., "New real-time closed-captioning system for Japanese broadcast news programs," *ICCHP*, pp. 651–654, July 2008.

[2] S. Homma, A. Kobayashi, T. Oku, S. Sato, and T. Imai, "Live Closed-Captioning with Robust Speech Recognition for a Spontaneous-Spoken Style," *APSIPA ASC*, pp. 450-453, December 2010.

[3] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, No. 9, pp. 171-185, 1995.

[4] J. -L. Gauvain and C. -H. Lee, "Maximum a Posterior Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech and Audio Process.*, Vol. 2, No. 2, pp. 291-298, April 1994.

[5] G. Saon and H. Soltau, "Boosting systems for large vocabulary continuous speech recognition," *Speech Communication*, Vol. 54, pp. 212-218, February 2012.

[6] H. Tang, M. Hasegawa-Johnson, and T. Huang, "Toward robust learning of the Gaussian mixture state emission densities for hidden Markov models," *ICASSP*, pp. 2274–2277, March 2010.

[7] V. V. Digalakis, and L. G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech and Audio Process.*, Vol. 4, No. 4, pp. 294-300, July, 1996

[8] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, Vol. 41, pp. 164–171, 1970.

[9] T. Imai, A. Kobayashi, S. Sato, H. Tanaka, and A. Ando, "Progressive 2-pass decoder for real-time broadcast news captioning," *ICASSP*, pp. 1937–1940, June 2000.