Fast Speaker Normalization and Adaptation based on BIC for Meeting Speech Recognition

Masato Mimura and Tatsuya Kawahara*
* Kyoto University, Academic Center for Computing and Media Studies,
Sakyo-ku, Kyoto 606-8501, Japan

***Abstract*—This paper presents a unified method for speech segmentation, speaker normalization of spectral features, and speaker adaptation of acoustic model for efficient meeting speech recognition. In the proposed method, input speech is segmented based on BIC (Bayesian Information Criterion), and compared against each speaker's statistic in the training corpus of the acoustic model based on the BIC. Fast VTLN (Vocal Tract Length Normalization) and MLLR (Maximum Likelihood Linear Regression) adaptation are realized using a pre-estimated warping factor and MLLR transformation matrices of the best-matched speakers, respectively. Experimental evaluations in Parliamentary speech transcription demonstrated that the proposed method achieved comparable ASR accuracy to the standard ML estimation for both VTLN and MLLR adaptation, with significant reduction of processing time.**

I. Introduction

In recent years, much research effort has been focused on the area of meeting speech transcription [1][2]. One of the challenging issues of meeting speech is that it contains many speakers and they appear by turns in an audio stream. For accurate speech recognition for a variety of speakers, speaker adaptation is essential for speaker-independent systems, and in order to apply it to meeting speech, the decoding process involves two extra stages, that is, segmentation of the audio data into speaker homogeneous segments and online estimation of adaptation parameters for each speaker turn. For speaker segmentation, a method based on BIC (Bayesian Information Criterion) is widely used because of its threshold-free characteristics and robustness [3].

Speaker adaptation has two major directions: one is speaker normalization of acoustic features, and the other is speaker adaptation of acoustic model. It is well-known that applying both feature normalization and model adaptation at the same time is effective. One of the most popular techniques for feature normalization is VTLN (Vocal Tract Length Normalization) [4] and for model adaptation, MLLR (Maximum Likelihood Linear Regression) [5] is widely used. Parameters for normalization and adaptation are typically calculated in an unsupervised manner, which involves further two separate stages: label generation and ML (Maximum Likelihood) estimation using the hypothesized label. In particular for speech recognition of meetings, in which speaker changes occur frequently, it is costly to estimate parameters for every speaker turn. While identification of speaker characteristics is a shared goal in these processes, the estimation procedures are usually separately designed and conducted in VTLN and MLLR, and also separated from the speaker segmentation process.

In this paper, we propose a unified method for speech segmentation, speaker normalization and adaptation based on the BIC for efficient meeting speech recognition. In the proposed method, the BIC used in speaker segmentation is used for identification of speaker characteristics, that is looking up similar speakers in the training speech corpus. Then, parameters for VTLN and MLLR of the corresponding speakers, which can be pre-computed offline, are loaded. Thus, we realize fast VTLN and MLLR adaptation, which do not need the label generation and ML estimation processes.

In the remainder of the paper, following brief introduction of the BIC in Section 2, the proposed method is described in Section 3. Then, evaluations of the method on Parliamentary speech transcription are presented in Section 4. Section 5 concludes the paper.

II. Speaker segmentation and clustering based on BIC

A. Speaker segmentation

Bayesian Information Criterion (BIC) [6] is a criterion for model selection. Given a data set $\{D_1, D_2, ..., D_N\}$ and candidate models M_1, M_2, ..., M_M, the BIC value of the model M_i is defined as

$$BIC(M_i) = \log P(D_1, D_2, ..., D_N | M_i) - \frac{1}{2}\lambda d_i \log N \qquad (1)$$

where d_i is the number of free parameters in the model M_i, and P is the likelihood of M_i for the data set. Then, M_i having the largest BIC value is selected to be the best model.

In BIC segmentation [3][7], given two consecutive segments S_1 (N_{S_1} samples) and S_2 (N_{S_2} samples), we consider two models: one representing the two segments with a single Gaussian $M_{S_1+S_2} = N(\mu_{S_1+S_2}, \Sigma_{S_1+S_2})$, and the other representing the two with respective Gaussians $M_{S_1,S_2} = \{M_{S_1}, M_{S_2}\} = \{N(\mu_{S_1}, \Sigma_{S_1}), N(\mu_{S_2}, \Sigma_{S_2})\}$. Then, BIC values for these two models are computed and compared. A full-covariance Gaussian distribution is usually assumed for each.

Specifically, $BIC(M_{S_1+S_2})$ is derived as:

$$
\begin{aligned}
BIC(M_{S_1+S_2}) &= -\frac{d}{2}(N_{S_1}+N_{S_2})\log 2\pi - \frac{N_{S_1}+N_{S_2}}{2} \\
&- \frac{N_{S_1}+N_{S_2}}{2}\log|\Sigma_{S_1+S_2}| \\
&- \frac{1}{2}\lambda(d+\frac{1}{2}d(d+1))\log(N_{S_1}+N_{S_2})
\end{aligned} \tag{2}
$$

where d is the dimension of the feature vector.

The model M_{S_1,S_2} has twice as many parametes as $M_{S_1+S_2}$, and its BIC becomes:

$$
\begin{aligned}
BIC(M_{S_1,S_2}) &= -\frac{d}{2}(N_{S_1}+N_{S_2})\log 2\pi - \frac{N_{S_1}+N_{S_2}}{2} \\
&- \frac{N_{S_1}}{2}\log|\Sigma_{S_1}| - \frac{N_{S_2}}{2}\log|\Sigma_{S_2}| \\
&- \lambda(d+\frac{1}{2}d(d+1))\log(N_{S_1}+N_{S_2})
\end{aligned} \tag{3}
$$

Then, the difference of these two, ΔBIC, is derived as:

$$
\begin{aligned}
\Delta BIC &= \frac{1}{2}((N_{S_1}+N_{S_2})\log|\Sigma_{S_1+S_2}| \\
&- N_{S_1}\log|\Sigma_{S_1}| - N_{S_2}\log|\Sigma_{S_2}|) \\
&- \frac{1}{2}\lambda(d+\frac{1}{2}d(d+1))\log(N_{S_1}+N_{S_2})
\end{aligned} \tag{4}
$$

Here λ is often called a penalty weight, and it is fixed to be 2.0 in this work.

If ΔBIC takes a positive value for the two consecutive segments, we can determine they should be modeled by different Gaussians, concluding that there is a changing point of some acoustic conditions (speaker change in meetings) between them.

Each segment identified by this procedure is considered to be acoustically homogeneous, therefore we can perform speaker normalization of acoustic features and speaker adaptation of acoustic model based on this segment. This scheme is widely used in broadcast news segmentation and recognition[8].

B. Speaker clustering and classification

In addition to segmentation, we can also perform clustering of speech segments based on the BIC. If ΔBIC for two segments (in particular, the current segment and any previous segment in the same meeting) takes a negative value, they are likely to originate from the same speaker cluster. Based on this clustering information, we can use shared parameters for the segments of the same speaker cluster. As a result, the number of online unsupervised parameter estimations can be reduced, and its reliability would be enhanced[1].

Furthermore, based on the BIC we can compare an input speech segment against not only a previous segment in the same meeting, but also a segment or a speaker cluster in the large speech corpus used for acoustic model training. In this work, we regard each speaker as a cluster by merging all his/her turn segments. For each speaker in the training corpus, we can estimate offline the parameters for speaker normalization and adaptation, using manual transcripts and an accurate context-dependent acoustic model, prior to the actual recognition. Thus, we do not need any online unsupervised estimation process (label generation and ML estimation), and fast normalization and adaptation can be realized using the pre-computed parameters.

[1]The effect was confirmed in preliminary experiments, though not presented in this paper.

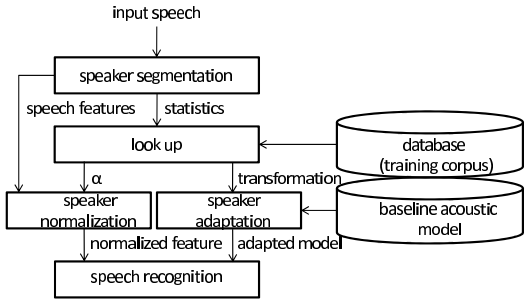

Fig. 1. Overview of the proposed method

There have been a number of previous works, which either search for the same speaker online (e.g. [9]) or search for similar speakers in the training corpus (e.g. [10]) to adapt/retrain the acoustic model. Our proposed method uses the same BIC used in speaker segmentation for searching for similar speakers, whose parameters are then used for speaker normalization and adaptation. Figure 1 illustrates the overview of the proposed method.

III. Fast speaker normalization and adaptation

A. Fast VTLN

We use VTLN [4] as a speaker normalization technique. VTLN is to normalize the spectral variation caused by different lengths of the vocal tract, and can be simply implemented by warping the frequency axis in the filterbank analysis. The warping is performed by a piece-wise linear function which is controlled by a scalar value: a warping factor α. The value of α for each input speech segment is estimated in an unsupervised manner during speech recognition. A typical procedure involves two stages; recognizing the segment with $\alpha = 1.0$ (no warping) and then performing forced alignment of the hypothesis for all possible warping factors (usually $0.8 \leq \alpha \leq 1.2$), to find the most likely value.

There are several techniques for more efficient implementation which do not require the initial transcription (ASR). One of the most popular techniques for fast VTLN is to use GMM for the warping factor selection [11]. In this method, GMM for each value of α is trained using speech data that will be processed with the same α value, and the model (corresponding value of α) that gives the highest likelihood to the unnormalized feature of the input speech is selected. Another

method is proposed by Emori et al. [12], which estimates α analytically and approximately based on the likelihood given by one specific acoustic model, without any search procedure for α, on the assumption that the formant position of an unknown speaker is not so far different from known speakers.

Our proposed method does not need any form of likelihood computation which is usually costly, but only looks up a same or very similar speaker in the database, and use the value of α of the speaker. The similarity is measured via ΔBIC which is used in speaker segmentation, and thus computed very efficiently.

B. Fast MLLR adaptation

We use MLLR [5], that is one of the most widely-used techniques for acoustic model adaptation. MLLR estimates a set of linear transformations for Gaussian means of HMM by ML criterion, instead of estimating the HMM parameters directly. These transformations shift the component means of the baseline model so that each state in the HMM is more likely to generate the adaptation data. MLLR works robustly with a smaller amount of adaptation data than direct adaptation such as MAP, because of the smaller number of the parameters. Unsupervised adaptation is performed in a similar manner as in VTLN; recognizing the adaptation data with the unadapted baseline model and then estimating the linear transformations with the hypothesized phone sequence.

Our proposed method does not conduct the initial transcription, but looks up a set of similar speakers in the database and use their transformation matrices. Unlike VTLN, which needs a single parameter, MLLR transformation involves a set of matrices with a number of parameters. Using a single speaker's information may not be reliable, unless he/she is the exactly same speaker. Thus, we use multiple (N-best) speakers' information to estimate the parameters for the input speech. In this work, we simply compute an average over the selected speakers. More sophisticated combinations such as a weighted mean should be explored in the future.

C. Fast comparison between input segment and training data via BIC

The search for similar speakers via ΔBIC can be done in a very efficient manner. In formula (4), assuming S_1 to be a current input speech segment and S_2 to be a speaker cluster in the training speech corpus, the second term $\log(|\Sigma_{S_1}|)$ has already been calculated during the segmentation process, and the third term $\log(|\Sigma_{S_2}|)$ were calculated and stored offline before the actual recognition. For calculating the first term, we need the covariance matrix of all samples in S_1 and S_2, but it can be calculated easily using the sufficient statistics of the two segments (the first order statistics, the second order statistics, and the number of frames), without access to each speech sample. These statistics should have been computed either offline or during the segmentation process.

IV. Experimental evaluations

Evaluation experiments were performed on speech transcription of the Japanese Parliament (Diet). We collected three committee meetings (Financial 7 hours, Judicial 2.5 hours, Education 2.5 hours, 12 hours in total) held in February 2010 for the test set. The total number of speakers was 54. Experiments were performed using both manual and automatic segmentation (automatic segmentation is based on ΔBIC, as described in Section 2). The total number of segments was 974 by manual segmentation, and 786 by automatic segmentation. For automatic segmentation and database look-up via ΔBIC, 24-dimensional unnormalized cepstrul features consisting of 12 MFCCs plus their Δ were used. On the other hand, 38-dimensional features consisting of 12 MFCCs, thier Δ, $\Delta\Delta$, and Δpower plus $\Delta\Delta$power were used for speech recognition.

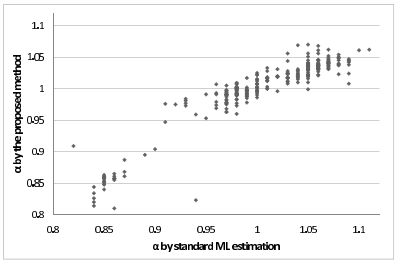

Fig. 2. Correlation of α estimated by the standard method and the proposed method

The acoustic model is triphone HMMs with 3000 states and 16 Gaussian components per state, trained by MPE criterion using meeting speech data (years 2001-2007). The amount of training data was 225 hours. Turn-based CMN (cepstrum mean normalization), CVN (cepstrum variance normalization) and VTLN were applied to the features.

The language model was built by applying statistical speaking-style transformation to ten-year (1999-2009) data of official meeting records. The text size was 170M words. The vocabulary size is 64K. Julius-4.1.2 was used for decoding.

The database for the proposed method was constructed using speaker turns in the training corpus that were longer than 30 seconds (9645 turns, 1189 speakers). The time interval between the training data and the test set is more than three years and 60% of the speakers in the test set were not present in the training corpus.

We preliminarily evaluated speaker recognition performance for 198 turns whose speakers have more than 10 minutes data in the training corpus. Speaker recognition was performed by choosing speakers in the training corpus who have the smallest ΔBIC values. The 1-best accuracy was only 33.9%, but the 5-best accuracy was 80.8%, and the 20-best accuracy was 89.7%. The accuracy with this simple method is not so high, but we expect similar speakers having a similar α value are selected.

The values of α estimated by the standard ML method and the proposed method for 287 turns of new speakers are plotted in Figure 2. Their correlation is 0.934, and the average error (difference of the two) is 0.017, showing that our method can estimate accurate warping factors without any transcription, even if the speakers are not present in the training corpus.

A. Evaluation of fast VTLN

We first evaluated the proposed fast VTLN. The standard method using the unsupervised (blind) ML estimation and the GMM-based selection method were also tested for comparison. A monophone acoustic model with 16 Gaussian components per state was used for label generation and forced alignment in the standard method. For the GMM-based estimation, we trained 41 GMMs for each value of α in the range $0.8 \leq \alpha \leq 1.2$ (step size of 0.01) using the corresponding speech data in the training corpus. The number of Gaussian components in each GMM is 16. The proposed method is much faster than the GMM-based method which needs to calculate the likelihood of the input speech frames. The proposed method took only 165 seconds of CPU time (CPU: Intel Xeon 3.0GHz) in estimating α for 12 hours of the test data, and was nearly 10 times faster than the GMM-based method which spent 1609 seconds.

The character accuracy of ASR resluts is summarized in Table I. In the case of manual segmentation, VTLN using the standard ML estimation gives improvement of 2.0% absolute compared to the result without VTLN. While the GMM-based method shows degradation of absolute 0.6% from the standard method, the proposed method achieves comparable ASR accuracy to the standard method. Much the same tendency is observed in the case of automatic segmentation.

B. Evaluation of fast MLLR adaptation

Then, we evaluated the proposed method for fast MLLR adaptation. The results are shown in Table II. For the proposed method, we show the results for two different numbers of N-best speakers used for adaptation, $N = 1$ and $N = 20$.

In the both cases of manual and automatic segmentation, MLLR using the proposed method shows further improvements (absolute 0.43-0.46$\%$) over the result with the proposed VTLN when $N = 20$, which are statistically significant at the 1% level. The degradations (0.25-0.28$\%$) from the standard MLLR are not statistically significant at the same level. On the other hand, the accuracy is degraded when the model is adapted with only 1-best speaker's transformation. The result coincides the fact that speaker recognition performance was low in the 1-best and needed the 20-best for a reliable level.

The standard method spent 19 hours of CPU time for the label generation and 1.3 hours for the ML parameter estimation for the entire test data, while the proposed method needed little additional time over the VTLN parameter estimation. Thus, the effect of speeding up is much more significant than the case of VTLN.

TABLE I
EFFECT OF FAST VTLN (CHARACTER ACCURACY $\%$)

method	manual segment	automatic segment
(without VTLN)	83.80	83.58
ML (blind search)	85.79	85.66
GMM-based selection	85.22	85.06
proposed method	85.63	85.54

TABLE II
EFFECT OF FAST MLLR (CHARACTER ACCURACY $\%$)

method	manual segment	automatic segment
proposed VTLN	85.63	85.54
proposed MLLR (N = 1)	85.22	85.08
proposed MLLR (N = 20)	86.06	86.00
standard MLLR	86.34	86.25

V. CONCLUSION

We have proposed an efficient method for speech segmentation, speaker normalization of the spectral features, and speaker adaptation of the acoustic model. The method consistently uses the BIC used in speech segmentation, for looking up similar speakers in the training corpus, whose parameters are then used for feature normalization and model adaptation. The proposed method achieved significant reduction of processing time without degradation of accuracy for meetings consisting of many speakers who are not covered in the training corpus.

This method is particularly useful for real-time captioning using speech recognision. As the proposed method is based on a very simple principle, it can be applied to other adaptation techniques such as Constrained MLLR and variance MLLR.

VI. ACKNOWLEDGEMENTS

This work was supported by JST CREST and JSPS Grant-in-Aid for Scientific Research.

REFERENCES

[1] S.Renals, T.Hain, and H.Bourlard, "Recognition and understanding of meetings: The AMI and AMIDA projects," in *Proc. IEEE Workshop Automatic Speech Recognition & Understanding*, 2007.

[2] N.Mirghafori, A.Stolcke, C.Wooters, T.Pirinen, I.Bulyko, D.Gelbart, M.Graciarena, S.Otterson, B.Peskin, and M.Ostendorf, "From Switchboard to Meetings: Development of the 2004 ICSI-SRI-UW Meeting Recognition System," in *Proc. ICSLP*, vol. III, 2004, pp. 1957–1960.

[3] S.Chen and P.Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127–132.

[4] L.Lee and R.C.Rose, "Speaker normalization using efficient frequency warping procedures," in *ICASSP*, 1996, pp. 353–356.

[5] C.J.Leggetter and P.C.Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," in *Computer Speech and Language*, vol. 9, 1995, pp. 171–185.

[6] G.Schwarz, "Estimating the dimension of a model," in *The Annals of Statistics*, vol. 6, 1978, pp. 461–464.

[7] M.Gettolo, M.Vescovi, and R.Rizzi, "Evaluation of bic-based algorithms for audio segmentation," in *Computer Speech and Language*, vol. 19.

[8] J.Gauvain, L.Lamel, and G.Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[9] Z.-P. Zhang, S. Furui, and K. Ohtsuki, "On-line incremental speaker adaptation with automatic speaker change detection," in *Proc. ICASSP*, vol. II, 2000, pp. 961–964.

[10] R.Gomez, T.Toda, H.Saruwatari, and K.Shikano, "Improving rapid unsupervised speaker adaptation based on hmm sufficient statistics," in *ICASSP*, vol. 1, 2006, pp. 1001–1004.

[11] L.Welling, S.Kanthak, and H.Ney, "Improved methods for vocal tract normalization," in *ICASSP*, 1999, pp. 761–764.

[12] T.Emori and K.Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation," in *EUROSPEECH*, 2001, pp. 1649–1652.