LOW-LATENCY SPEAKER DIARIZATION BASED ON BAYESIAN INFORMATION CRITERION WITH MULTIPLE PHONEME CLASSES

Takahiro Oku, Shoei Sato, Akio Kobayashi, Shinichi Homma, and Toru Imai

NHK (Japan Broadcasting Corporation)
Science & Technology Research Laboratories, Tokyo, Japan

ABSTRACT

Low-latency speaker diarization is desirable for online-oriented speaker adaptation in real-time speech recognition. Especially in spontaneous conversations, several speakers tend to speak alternatively and continuously without any silence in between utterances. We therefore propose a speaker diarization method that detects speaker-change points and determines the speaker with a fixed low latency on the basis of a Bayesian information criterion (BIC) by using acoustic features classified into multiple phoneme classes. To improve the accuracy of speaker diarization in the low latency condition, the speaker-decision is made continuously at each phoneme boundary. In an experiment on conversational broadcast news programs, our diarization method reduced the speaker diarization error rate relatively by 20.0% compared to the conventional BIC with a single phoneme class. The online speaker adaptation applied in a speech-recognition experiment reduced word error rate at speaker-change points relatively by 7.8%.

Index Terms— speaker diarization, BIC, phoneme classes, speaker adaptation, speech recognition

1. INTRODUCTION

A low-latency speaker diarization for spontaneous and conversational speech recognition is proposed. If the question "who is speaking now" were accurately resolved with a low latency from a given speech, the performance of real-time speech recognition would be improved by techniques such as speaker adaptation [1].

Speaker-diarization systems are divided into two types: online or offline. In the case of general offline types, an audio stream is first split into several smaller segments that are then agglomerated according to a criterion such as the Bayesian information criterion (BIC) or the generalized likelihood ratio (GLR). On the other hand, online systems determine whether a speech segment belongs to one of the known speakers or a new speaker without use of future speech data. For real-time speech recognition during live TV programs, which is our target application, online systems should shorten the processing delay.

A conventional method for online speaker diarization consists of two steps. First, speech segments are extracted from an audio stream using energy-based or model-based voice activity detection (VAD). Second, the VAD-based segments are clustered and labeled with speaker labels. The accuracy of this diarization method, however, may degrade under spontaneous and conversational conditions, since continuous utterances by multiple speakers may not be properly segmented and the VAD-based segment consisting of multiple speakers may be identically clustered. In this case, subsequent audio segments cannot be clustered successfully either. Liu, et al. proposed a method that detects points at which the speaker changes (hereafter, "speaker-change points") with respect to each phoneme boundary for extracting the segments, each of which belongs to only one speaker [2, 3]. However, the speaker clustering is performed at the end of the speaker's turn, so the speaker-decision latency becomes high. A low-latency diarization system has been proposed [4], but it clusters the VAD-based segments and does not take speaker changes inside one segment into consideration.

In general speaker recognition or identification, speaker models are often expressed by Gaussian-mixture models (GMMs) [1, 4]. However, in the tasks of speaker-change detection and speaker-decision with low-latency, very short segments, which result in less reliable statistics for the GMMs, must be dealt with.

A speaker-diarization system—which performs speaker-change detection and speaker clustering based on BIC with multiple phoneme classes—for accurately detecting speaker differences even in a short segment is proposed in the following. A perceptual evaluation of speaker identification reported vowels and nasals are effective for distinguishing speakers [5]. Classification of the acoustic features in accordance with phoneme information is therefore expected to have a beneficial effect on speaker diarization, especially in a low-latency situation. In this study, to improve the accuracy of speaker diarization for real-time speech recognition, speaker models with these classified features were used.

Furthermore, a speaker-adaptation method for speech recognition using the speaker labels given from the speaker-diarization running in parallel is proposed. Our speech recognition switches acoustic models on the basis of the speaker labels even inside a speech segment. The acoustic model is adapted by the maximum likelihood linear regression (MLLR) at every speaker-change point, where the MLLR uses the speech data before the change points.

This paper is organized as follows. Section 2 describes the proposed online speaker diarization (based on the BIC with multiple phoneme classes) and the online speaker adaptation for real-time speech recognition. Section 3 describes the experimental setup for performing speaker diarization and speaker adaptation on conversational programs in Japanese broadcasting and presents the experimental results. Section 4 concludes the study and mentions future works.

2. ONLINE SPEAKER DIARIZATION AND ADAPTATION

2.1. System overview

The proposed system for speaker diarization and speech recognition is shown schematically in Fig. 1. Using the same acoustic features with speech recognition, the speaker diarization de-

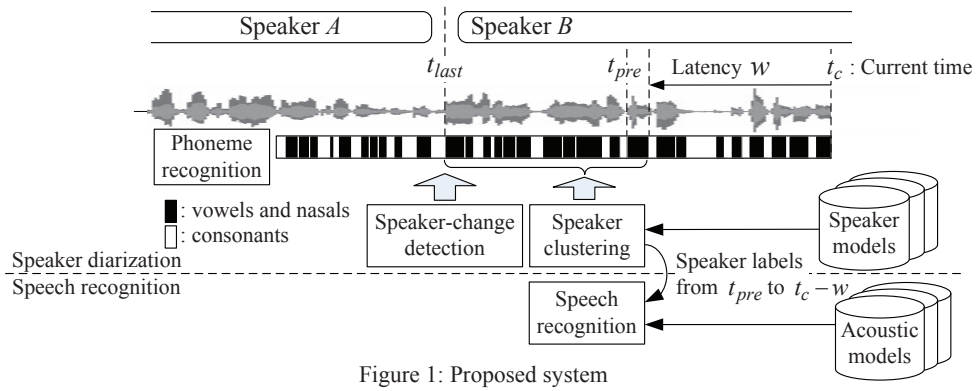

Figure 1: Proposed system

termines speakers in real time. The acoustic features are 39 dimensional ones including 12 mel-frequency cepstral coefficients (MFCCs), log-power, and their first- and second-order time derivatives. In speaker diarization, the acoustic features are classified into a vowel class, which also includes nasals in this paper, and a consonant class in accordance with phoneme information. The vowel class is expected to contain more speaker's acoustic individuality than a single class with all phonemes. The phoneme information is extracted by real-time phoneme recognition [6]. The proposed method detects speaker changes and determines a speaker on the basis of BIC with the multiple phoneme classes with a fixed latency. No speaker model is necessary at the start of the speaker diarization. The speaker models are created or updated at each speaker-change point by using all the speech assigned to the speaker before the change point. It is possible to set pre-trained speaker models for specific types of speakers, such as known anchorpersons.

Speech recognition is performed while the system switches among acoustic models in accordance with a speaker label assigned through the speaker clustering. The acoustic models used for the recognition are also adapted at each speaker-change point in the same manner as the speaker models. The speaker adaptation is performed by MLLR using the speech before the speaker change and the corresponding recognition result.

2.2. BIC with multiple phoneme classes

Online speaker diarization comprises two procedures: speaker-change detection and speaker clustering. ΔBIC based on BIC [7] is used in both procedures. Widely used to diarize speakers, ΔBIC is a criterion for determining whether sets of feature vectors x and y come from the same speaker or from two distinct speakers. It is given as

$$\begin{aligned}\Delta BIC(x,y) &= \log\frac{p(x\,|\,\lambda_x)\cdot p(y\,|\,\lambda_y)}{p(xy\,|\,\lambda_{xy})} - \alpha P(f_{xy}, d) \\ &= \frac{1}{2}\left[f_{xy}\log|\Sigma_{xy}| - f_x\log|\Sigma_x| - f_y\log|\Sigma_y|\right] - \alpha\left(\frac{d(d+3)}{4}\right)\log(f_{xy}).\end{aligned} \quad (1)$$

Here, λ_x is a speaker model, which is a single Gaussian model with covariance matrix Σ_x ML-estimated on the corresponding data x, P is a penalty factor composed of a number of frames f and dimensions d of a feature vector, and α is a penalty weight. xy represents a concatenation of the vector sets of x and y. f_{xy} is therefore the sum of f_x and f_y, and λ_{xy} represents a speaker model in which x and y are derived from the same speaker. When ΔBIC is greater than zero, x and y are regarded as segments uttered by distinct speakers; otherwise, they are regarded as being from the same speaker.

In [8], Eq. (1) is extended to mixture models as follows:

$$\begin{aligned}\Delta BIC(x,y) &= \log\left(\frac{\prod_{m=1}^{M} p(x_m|\lambda_x^m)\,p(y_m|\lambda_y^m)}{\prod_{m=1}^{M} p(x_m\,y_m|\lambda_{xy}^m)}\right) - \alpha P \\ &= \frac{1}{2}\left[\sum_{m=1}^{M} f_{xy}^m \log|\Sigma_{xy}^m| - \sum_{m=1}^{M} f_x^m \log|\Sigma_x^m| - \sum_{m=1}^{M} f_y^m \log|\Sigma_y^m|\right] \\ &\quad - \alpha M\left(\frac{d(d+3)}{4}\right)\log(f_{xy}),\end{aligned} \quad (2)$$

where M is a number of mixtures, and λ_x^m represents the model estimated on x assigned to the m-th mixture distribution considering hard alignment of frames to the mixture components.

To express a speaker model with multiple phoneme classes, we regard M as a number of distinctive phoneme classes instead of mixture models, and λ_x^m represents the model estimated on x classified into the m-th phoneme class. The proposed diarization is implemented using ΔBIC given by Eq. (2), and M is replaced with 2 for the vowel class and the consonant class, unlike the conventional ΔBIC (with a single phoneme class) indicated by Eq. (1).

2.3. Speaker change detection

The proposed speaker-change detection restricts the change points at phoneme boundaries only [2]. A speaker-change point is detected from a collection of phoneme boundaries, $T_{hyp} = \{t_{last}, \cdots, t_c\}$, where t_{last} and t_c correspond to the previously detected speaker-change point and current time, respectively. T_{hyp} is extracted by phoneme recognition [6] for speech detection. The phoneme boundary t_h that satisfies equation (3) and inequation (4) is regarded as a speaker-change point.

$$t_h = \arg\max_{t_k \in T_{hyp}} \Delta BIC\left(x(t_{last} : t_k], x(t_k : t_c]\right) \quad (3)$$

$$\Delta BIC\left(x(t_{last} : t_h], x(t_h : t_c]\right) > 0 \quad (4)$$

Here, $x(t : t']$ stands for the acoustic features between $t+1$ and t'. As soon as a speaker-change point is detected, the following

speaker clustering determines who is the speaker (with a fixed low-latency) as described in the next section.

2.4. Speaker clustering

In conventional online speaker diarization, speaker-decision is made at the end of the speaker's turn or VAD-based segment end; consequently, the decision latency becomes high. Accordingly, we propose a new speaker-decision method with a low-latency w, which is a predetermined and fixed value, while continuously making the decision at every phoneme boundary for more accurate speaker labels with more speech data. This proposed decision method is illustrated schematically in Fig. 1 in the case that two people, A and B, have a conversation, and the speaker continuously changes between A to B at t_{last}. Their continuous utterances may not be properly segmented on the basis of VAD. With the proposed method, at current time t_c, the speaker from t_{last} to $t_c - w$ is estimated by using the statistics from t_{last} to t_c. It is carried out at every phoneme boundary, because the estimated speaker especially shortly after the speaker-change points is unstable. Since a speaker label given at the previous speaker-decision time t_{pre} is already used in speech recognition, the latest speaker label given at $t_c - w$ is used as a final one from t_{pre} to $t_c - w$, and t_{pre} is updated and replaced by $t_c - w$ for the next speaker-decision.

The criterion used for clustering the speaker is

$$\Delta BIC_i\left(\bar{x}_i, x\left(t_{last} : t_c\right]\right) \quad (i \in C), \tag{5}$$

where i is the speaker label belonging to a collection of speaker clusters C that is registered in the system, and $\bar{x}_i$ stands for feature vectors corresponding to speaker i. A new speaker is assigned from t_{last} to $t_c - w$ if $\Delta BIC_i > 0 \ \ \forall i \in C$. Otherwise, the speaker is determined to be $\hat{i}$ according to

$$\hat{i} = \arg\min_{i \in C} \Delta BIC_i\left(\bar{x}_i, x\left(t_{last} : t_c\right]\right). \tag{6}$$

Speaker model λ_{x_i} is expressed by frame numbers of multiple classes, $f_{x_i}^m$, and covariance matrices, $\Sigma_{x_i}^m$. λ_{x_i} is updated or a new speaker model is generated every time a new speaker-change point is detected by the speaker-change detection procedure using all the speech assigned to the speaker before the change point.

2.5. Speaker adaptation for speech recognition

Real-time speech recognition switches acoustic models to the right speaker's models at every phoneme boundary on the basis of the speaker labels sequentially provided by the speaker-decision procedure. If a speaker change is detected, the speech before the change point is used for the acoustic-model adaptation by the MLLR. The detailed procedure is as follows.

1) Speaker labels are continuously sent from the speaker diarization to the speech recognizer with a fixed latency w. It is supposed that a speaker change from one to another, for example, from speaker A to speaker B, is detected.

2) The acoustic model used for the speech recognition is switched from speaker A's model to speaker B's model with a fixed latency w.

3) The acoustic model of speaker A is adapted by using the speech and corresponding recognition results obtained before the speaker-change point. The adapted model is then used when speaker A speaks next.

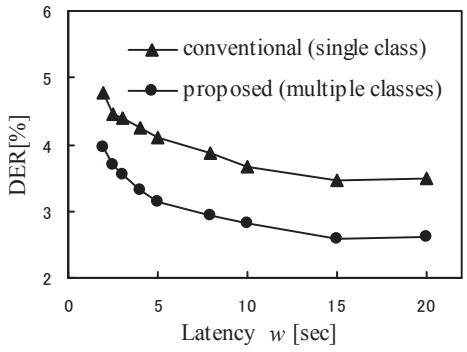

Figure 2: Comparison of DERs

3. EXPERIMENT

3.1. Evaluation set

The proposed system was evaluated using a Japanese TV talk show called "*Today's Close-up*," composed of a conversation between a newscaster and various guests in the studio. This evaluation set, obtained from seven episodes that aired in May 2008, comprised 12,356 words uttered by ten speakers with 120 speaker changes and a 3,177-second long speech.

3.2. Speaker diarization results

A penalty weight α for speaker-change detection and speaker clustering was determined by using episodes of the program aired in the week before the evaluation data. Diarization error rate (DER) was used as a diarization evaluation metric. DER is the time-weighted sum of false alarm speech (FS), missed speech (MS), and speaker error (SE). The acoustic features used in the proposed diarization system correspond to all phonemes (except silences) extracted by means of phoneme recognition. The accuracy of phoneme classification (vowel class and consonant class) for the evaluation set was 73%. FS and MS, which were determined by the phoneme recognition, were 0.6% and 0.4%, respectively. One female anchorperson's speaker model was trained from 31-hour speech data and assigned to the diarization system in advance. The other speaker models were sequentially created and assigned to the system during the online speaker diarization.

In this experiment, the proposed method using ΔBIC with multiple phoneme classes (vowel class and consonant class) was compared with a conventional diarization method using ΔBIC with a single phoneme class containing acoustic features for all the phonemes with different fixed latencies. It was also compared with a GLR-based method and a GMM-based method. In this comparison, speaker was determined at the end of the speaker's turn, and the decision latency was the length of audio speech between the speaker-change points. The GLR-based method replaced ΔBIC with GLR [2] for speaker-change detection and speaker clustering. The GMM-based method [4] with 128 Gaussian mixtures was employed only for speaker clustering with ΔBIC-based speaker-change detection.

Diarization results obtained by the proposed and conventional methods (with different latencies w from 2 to 20 seconds) are given in Fig. 2. The proposed diarization method reduced DERs

Table 1: Comparison of WERs [%]

Evaluation data	Non-adapted	Speaker-adapted						
		Oracle	Conventional (single class)			Proposed (multiple classes)		
			$w = 2$	$w = 5$	$w = 10$	$w = 2$	$w = 5$	$w = 10$
a) All	21.7	19.4	20.4	20.0	20.0	20.2	19.9	19.9
b) At speaker change points	24.5	21.0	**24.4**	22.5	22.3	**22.5**	22.3	22.1

compared with the conventional method for all w. DER was 4.0% with 2-second latency for the proposed method, while it was 4.8% for the conventional method. A diarization error reduction rate was 20.0%. The DERs for determining a speaker at the end of the speaker's turn for the proposed method, the GLR-based method, and the GMM-based method were 2.6%, 5.7%, and 12.6%, respectively. These results confirm that modeling speakers with multiple phoneme classes based on BIC is more effective than the other methods especially in a low latency.

3.2. Speaker adaptation results

The speaker adaptation for real-time speech recognition was performed by using the speaker labels from the speaker diarization. Original acoustic models to be adapted were gender-dependent models that were trained from NHK's Japanese broadcast news consisting of 340 hours of male and 250 hours of female utterances. As well as the speaker model for speaker diarization, an acoustic model of the anchorperson was also trained from 31-hour speech data in advance. This anchorperson's acoustic model was used for the recognition when the diarization system identified her speech, and was not adapted any more because her acoustic model was considered to be already trained by enough speech data. The other acoustic models for the guests were speaker-adapted during online speech recognition with speaker labels and speech-recognition results.

Table 1 compares the word error rates (WERs) for four conditions: non-adapted, oracle adaptation, adaptation based on diarization with a conventional single phoneme class, and adaptation based on the proposed diarization with multiple phoneme classes. In the non-adapted case, dual-gender speech recognition [6] was carried out by using the gender-dependent and speaker-independent acoustic models. Oracle adaptation means that the acoustic models were adapted with true speaker labels. As well as the results of the speaker diarization, the performance of the speaker adaptation based on the proposed diarization was slightly better than that of the conventional adaptation for all the evaluation data (Table 1-a). The WERs were reduced as latency w increased. With 2-second latency, the improvement of the speaker adaptation was less than that in the case with 5 or 10-second latency. This is because the 2-second latency brings about later speaker-change detection and delayed switching to the right acoustic models for the recognition. This switching delay degraded the recognition performance. Because the timing of the switching of the acoustic model is influenced by the accuracy of the speaker-diarization performance around the actual speaker changes, the previous or next sentences at speaker-change points were evaluated. As a result, the WERs of the conventional and proposed methods were 24.4% and 22.5%, respectively, for 2-second latency (word error reduction rate of 7.8%) at the speaker-change points (Table 1-b). There was significant difference between the conventional and the proposed methods using a matched-paired difference t-test at a significance level of 0.05. This result confirms that the proposed method improves the performance of speech recognition, especially around the speaker changes in a low latency.

4. CONCLUSION

A low-latency speaker-diarization system based on BIC with multiple phoneme classes for conversational speech was proposed. To improve the performance of the diarization, acoustic features are classified into multiple phoneme classes (vowel class and consonant class) by phoneme recognition. An online method for determining the speaker during the real-time speech recognition is proposed. Furthermore, an online speaker-adaptation method using speaker information extracted from the diarization system is proposed. Experiments were performed on conversational broadcast programs. Our proposed method reduces speaker-diarization error rate relatively by 20.0%. Applying the online speaker adaptation in the speech-recognition experiment reduced a word error rate relatively by 7.8% at speaker-change points. Future work will involve further improving the low-latency speaker diarization, especially right after the actual speaker changes. Since speech data are accumulated for speaker adaptation, improving diarization around the speaker changes is quite important for speech recognition with adapted acoustic models.

4. REFERENCES

[1] Z. Zhang, S. Furui, and K. Ohtsuki, "On-line Incremental Speaker Adaptation for Broadcast News Transcription," *Speech Communication*, Vol. 37, No. 3, pp. 271-281, July. 2002

[2] D. Liu and F. Kubala, "Fast Speaker Change Detection for Broadcast News Transcription and Indexing," *Proc. EUROSPEECH'99*, Vol. 3, pp. 1031–1034, 1999

[3] D. Liu and F. Kubala, "Online Speaker Clustering," *Proc. IEEE ICASSP*, pp. 333–336, 2004

[4] K. Markov and S. Nakamura, "Improved Novelty Detection for Online GMM Based Speaker Diarization," *Proc. INTERSPEECH*, pp. 363–366, 2008

[5] K. Amino and T. Arai, "Effect of Linguistic Contents on Perceptual Speaker Identification: Comparison of Familiar and Unknown Speaker Identifications," *Acoust. Sci. & Tech.*, Vol. 30, No. 2, pp. 89-99, 2009

[6] T. Imai, S. Sato, A. Kobayashi, K. Onoe, and S. Homma, "Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News," *Proc. INTERSPEECH*, pp. 1602–1605, 2006

[7] S. Chen and P. Gopalakrishnam, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132, 1998

[8] K. Mori and S. Nakagawa, "Speaker Change Detection and Speaker Clustering Using VQ Distortion for Broadcast News Speech Recognition," *Proc. IEEE ICASSP*, pp. 413-416, 2001