ADAPTIVE AND ONLINE SPEAKER DIARIZATION FOR MEETING DATA

Giovanni Soldi[1]*, Christophe Beaugeant*[2] *and Nicholas Evans*[1]

[1]Multimedia Communications Department, EURECOM, Sophia Antipolis, France
[2] Intel Mobile Communications, Sophia Antipolis, France,
[1] {soldi,evans}@eurecom.fr, [2]christophe.beaugeant@intel.com

ABSTRACT

Speaker diarization aims to determine 'who spoke when' in a given audio stream. Different applications, such as document structuring or information retrieval have led to the exploration of speaker diarization in many different domains, from broadcast news to lectures, phone conversations and meetings. Almost all current diarization systems are offline and ill-suited to the growing need for online or real-time diarization, stemming from the increasing popularity of powerful, mobile smart devices. While a small number of such systems have been reported, truly online diarization systems for challenging and highly spontaneous meeting data are lacking. This paper reports our work to develop an adaptive and online diarization system using the NIST Rich Transcription meetings corpora. While not dissimilar to those previously reported for less challenging domains, high diarization error rates illustrate the challenge ahead and lead to some ideas to improve performance through future research.

Index Terms— Speaker diarization, clustering and segmentation, online diarization

1. INTRODUCTION

Speaker diarization [1–3] is an unsupervised statistical pattern recognition task which aims to determine 'who spoke when' in a given audio stream. Speaker diarization has become a key, enabling technology in a wide variety of tasks including document processing, structuring and navigation, information retrieval, meta-data extraction and copyright detection. Data domains include broadcast news, lectures, phone conversations and meetings. The latter is generally considered to be the most challenging on account of the high level of speaker overlap, spontaneity and short speaker turns.

Historically, the state-of-the art in speaker diarization for meetings has evolved around the implementation of offline systems, such as bottom-up and top-down hierarchical clustering approaches [3–5]. In both cases, speakers are modelled with Gaussian mixture models (GMMs) which are interconnected to form an ergodic hidden Markov model (HMM) in which the transitions represent speaker turns. These approaches, usually coupled with other segmentation and clustering techniques are applied offline to an entire audio stream to generate a segmentation hypothesis which resembles the real number of speaker and speaker turns.

With the increasing popularity of powerful, mobile smart devices, there is now a growing interest to develop online speaker diarization systems. Due to their computational complexity and high latency, the existing state-of-art diarization techniques are not easily adapted to online processing. There is thus an interest to develop entirely new, online approaches. A small number have been reported previously, but the majority, e.g. [6–8], focus on applications involving plenary speeches and broadcast news. Inspired by the approach in [6], this paper presents our efforts to develop such a system for the online diarization of meetings.

The remainder of this paper is organized as follows. Section 2 outlines previous, related work. Section 3 describes the online diarization system. Section 4 describes the experimental setup used to obtain results presented in Section 5. Our conclusions are presented in Section 6.

2. PRIOR WORK

Driven by the increasing popularity of powerful, mobile smart devices and the need for real-time information extraction in diverse human interaction scenarios, online diarization has attracted increasing interest in recent years. Although real-time diarization can be performed efficiently with the aid of multiple microphones and cameras [9, 10], scenarios in which only a single microphone is available remain challenging. Online diarization performance is typically far from what can be achieved with offline approaches. This section presents a review of the past work.

Liu et al. [11] present an approach in the context of broadcast news diarization. Speech activity detection (SAD) is applied to identify speech segments which are clustered via one of two different algorithms in order to perform online diarization. The study, involving leader-follower, dispersion-based and combined clustering algorithms was conducted with the NIST Hub4 1996 broadcast news database.

Markov et al. [7, 8] investigated more conventional speaker diarization using GMMs. Non-speech segments are discarded using a suitably trained GMM whereas diarization is performed upon the comparison of speech segments to a set of speaker models. New speaker models are introduced using an incremental expectation-maximization (EM) algorithm. The system was assessed on a database of European Parliament plenary speeches for which a diarization error rate (DER) of 8% was reported.

A similar approach was reported by Geiger et al. [6] for broadcast news. Here, speaker models were learned through maximum-a-posteriori adaptation (MAP) of a universal backgroud model (UBM). The same UBM is used to control the attribution of speech segments to existing speakers and the addition of new speaker models. A DER of 39% was reported.

Vaquero et al. [12] present a hybrid system composed of offline diarization [5] and online speaker identification. An initial offline diarization stage is used to learn speaker models. An online speaker identification system is then used for subsequent diarization. Speaker model adaptation is performed in parallel. Performance is dependent on the latency and accuracy of the offline process. A DER of 38% is reported for a set of 26 meetings from the NIST Rich Transcription (RT) evaluation corpora.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

Oku et al. [13] report a low-latency, online speaker diarization system by exploiting phonetic information in order to estimate more discriminative speaker models. Phone boundaries are considered as potential speaker turns. Features are initially clustered into predefined acoustic classes. GMM speaker models have the same number of components as the number of acoustic classes. A traditional delta-BIC-like criterion is then used for speaker clustering and segmentation. Performance is assessed using Japanese TV talk shows where conversations are characterized by short speaker turns and only few silence intervals.

Although there is a growing body of work with a focus on online diarization, a truly online, usable system for meeting data is lacking. This paper presents our efforts to develop such a system using standard meeting data from the NIST RT (RT) evaluation corpora.

3. ONLINE DIARIZATION SYSTEM IMPLEMENTATION

The online diarization system is illustrated in Fig. 1. It is based on the top-down or divisive hierarchical clustering approach to offline diarization reported in [4] and the online diarization approach reported in [6]. Aside from background modelling, there are two stages: (i) feature extraction; (ii) speech activity detection and (iii) online classification.

3.1. Feature extraction

Audio files are initially treated with Wiener filtering noise reduction. The signal is then frame blocked, windowed and parametrised with Mel-frequency cepstral coefficients plus energy. Parameters are augmented with delta and acceleration coefficients to produce a series of acoustic observations $\mathbf{o}_1, \ldots, \mathbf{o}_T$. Critically, for any time $\tau \in 1, \ldots, T$ only those observations for $t < \tau$ are used for diarization.

3.2. Speech activity detection and online classification

Non-speech segments are removed according to the output of a conventional model-based speech activity detector (SAD). The remaining speech segments are then divided into smaller sub-segments whose duration is no longer than an a-priori fixed maximum duration T_S. Online classification is then applied in sequence to each segment.

Segments are either attributed to an existing speaker model, or a new speaker model is created. This procedure is controlled with a universal background model (UBM) denoted by s_0 which is trained on external data. New speaker models are introduced if the current segment i generates a higher log-likelihood when compared to the UBM than to a set of speaker models s_j, where $j = 1, \ldots, N$ and where N indicates the number of speakers in the current hypothesis. Segments are attributed according to:

$$s_j = \underset{l \in (0,\ldots,N)}{\arg\max} \sum_{k=1}^{K} \mathcal{L}\left(\mathbf{o}_k | s_l\right) \tag{1}$$

where $\mathbf{o}_k$ is the k-th acoustic feature in the segment i, K represents the number of acoustic features in the i-th segment and where $\mathcal{L}\left(\mathbf{o}_k | s_l\right)$ denotes the log-likelihood of the k-th feature in segment i given the GMM model s_l. If the segment is attributed to s_0 then a new speaker model s_{N+1} is learned by MAP adaptation of the UBM model s_0 using the features contained in segment i. The segment is then labelled according to the newly introduced speaker and N is increased by one. When a segment is attributed to an existing speaker, then the corresponding model is adapted through maximum a-posteriori (MAP) adaptation. The segment is then labelled according to the recognised speaker j as per Eq. 1.

4. EXPERIMENTAL SETUP

In contrast to previous work this paper is concerned with the online diarization of the most challenging meeting data amassed from the set of NIST RT corpora. The full experimental setup is described here.

4.1. Databases, feature extraction and UBM

Experiments were performed using three independent datasets:

1. **RTubm:** a set of 16 meeting shows from the NIST RT'04 evaluations;
2. **RTdev:** a set of 15 meeting shows from the RT'05 and RT'06 evaluations, and
3. **RTeval:** a set of 15 meeting shows from the RT'07 and RT'09 evaluations.

The average show duration within the RTubm, RTdev and RTeval sets is 10, 15 and 24 minutes respectively. The average number of speakers within each set is 5, 5 and 4 respectively. All experiments concern the most challenging, single distant microphone (SDM) condition. Each acoustic signal is characterized by 12 Mel-frequency cepstral coefficients (MFCCs) augmented by energy, delta and acceleration coefficients thereby obtaining feature vectors with a total of 39 coefficients and computed every 10ms using a 20ms window.

The UBM model s_0 is trained offline using data in the RTubm set. Experiments were performed with different model sizes: 8, 16, 32, 64 and 128 Gaussian components. In all cases, non-speech segments are removed according to ground-truth transcriptions; UBMs are trained using only the remaining speech segments.

4.2. Online diarization

The system was developed and optimised using only the RTdev set. The dependence of the segment duration and model size on online diarization performance is assessed and presented here independently for both the RTdev and RTeval sets; both sets are small and the acknowledged differences between them means that it is of interest to analyse performance on both sets. Experiments are reported for segment durations of $0.25, 0.5, 1, \ldots, 10$ seconds and model sizes of between 8 and 128. Speech activity detection is applied in the same way, with diarization being applied only to remaining speech segments.

4.3. Assessment and metrics

Diarization performance is assessed using the standard diarization error rate (DER) metric with the usual 0.25 second collar applied to each speaker turn; diarization errors within the collar are ignored. Performance is first assessed as a function of segment duration for differing model sizes. Dynamic convergence performance is then assessed periodically at each minute T_i. For maximum statistical significance, assessment is nonetheless performed on the entire show. While this approach to *assessment* might appear more in keeping with offline diarization, results still reflect online performance; it uses speaker models learned through the online process only up until minute T_i. While diarization performance should improve naturally

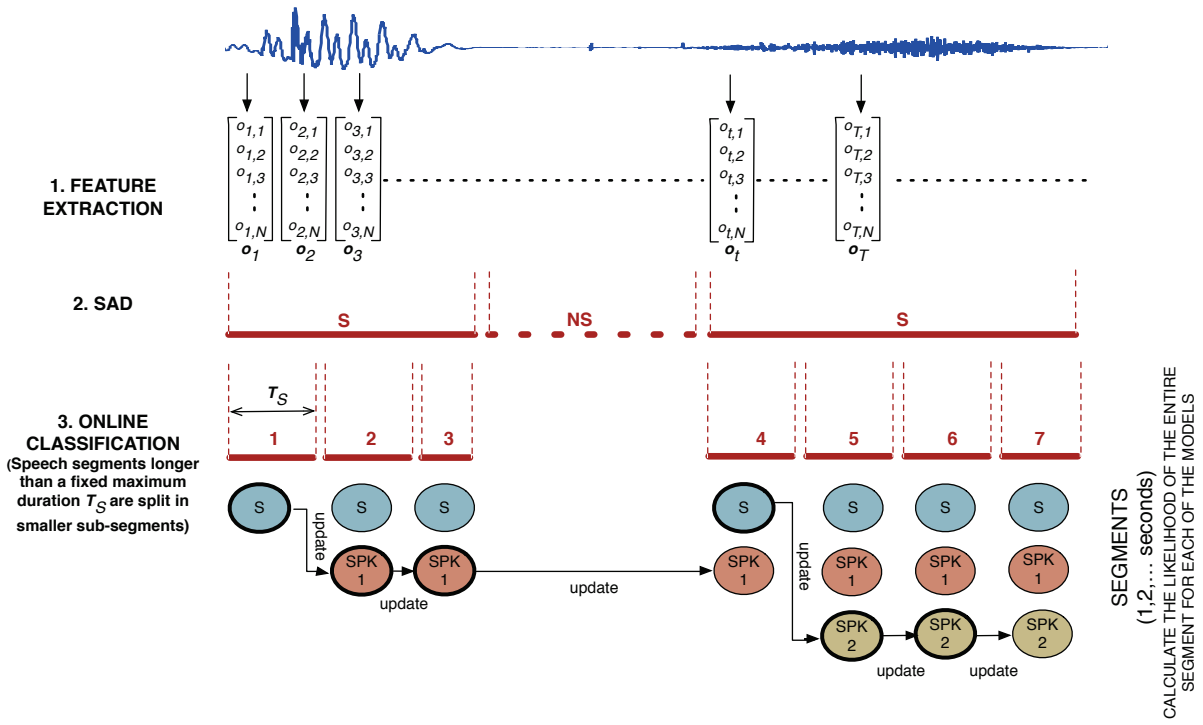

Fig. 1: An illustration of the online speaker diarization system.

as the full set of speakers is gradually introduced into the online process, as illustrated in Fig. 4 (red line) 90% of speakers appear in the first 2 to 3 minutes of each show. This approach to assessment is therefore still representative of online performance.

Each segment is attributed to one of the speaker models according to the highest likelihood criteria, without any model adaptation or re-segmentation. While the application of adaptation and re-segmentation would improve performance, they would also introduce latency not in keeping with online diarization. Adaptation and re-segmentation would furthermore add to the computational complexity and so we did not investigate their use.

5. EXPERIMENTAL RESULTS

DER results are assessed as a function of segment duration, model size and the amount of training data available as a function of time T_i. Also presented is an analysis of dynamic speaker statistics in addition to some ideas for future work.

5.1. Segment duration and model size

Figure 2 illustrates online diarization performance in terms of DER as a function of segment duration and model size. Plots are illustrated for the RTdev set (left) and RTeval set (right). The optimal model size is either 32 or 64 Gaussian components, with the larger model size being the most consistent across the two sets. Performance deteriorates as the model size is increased further and is due to a lack of sufficient data for reliable learning. Initially, the DER tends to decrease as the segment duration increases. As the segment size increases beyond the optimum, then more and more speaker turns are missed causing the DER to increase. Across the two datasets, the minimum DER is between 40% and 45%. This is a high error rate, but one not dissimilar to that reported in previous work performed using broadcast new data, e.g. [6].

5.2. Adaptive speaker modelling and convergence

Figure 3 illustrates the dynamic convergence of the DER as a function of time T_i. Plots are again illustrated for the RTdev set (left) and RTeval set (right) and for segment durations of between 1 and 6 seconds. All plots are for GMMs with 64 components. Both figures show that the DER decreases as the amount of data available for model training increases. The plots also show that segment durations of less than 2 seconds are largely insufficient for reliable diarization, with optimal performance being achieved with 2 or 3 second segments. Again, as the segment size increases, then more and more mid-segment speaker turns are missed, thus leading to higher DERs.

5.3. Dynamic speaker statistics

Figure 4 illustrates the evolution in speaker numbers for the RTdev set. Profiles are shown for the ground truth references (red line) and the corresponding number appearing in the automatically generated diarization hypothesis (blue line). The hypothesis corresponding to GMMs of 64 components and segments length of 2 seconds. For the first five to six minutes, the hypothesis contains fewer speakers than the ground truth whereas, beyond, the hypothesis contains more speakers than the ground truth.

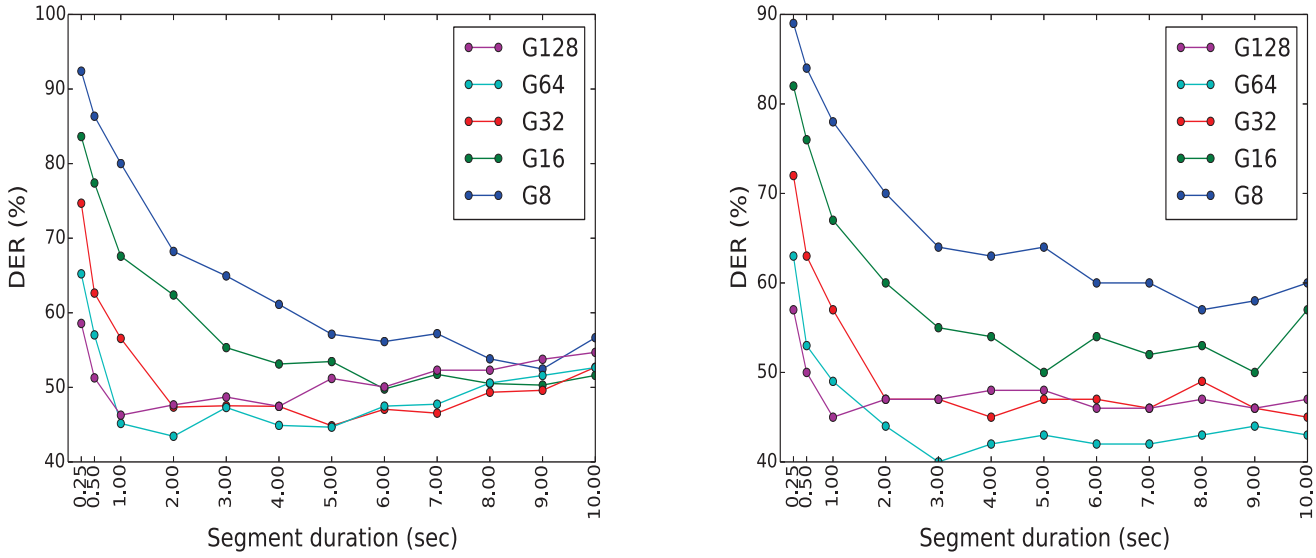

Fig. 2: An illustration of DER as a function of segment duration (0.25,0.5,1-10 sec) and for different model sizes (8-128). Results shown for the RTdev (left) and RTeval (right) datasets.

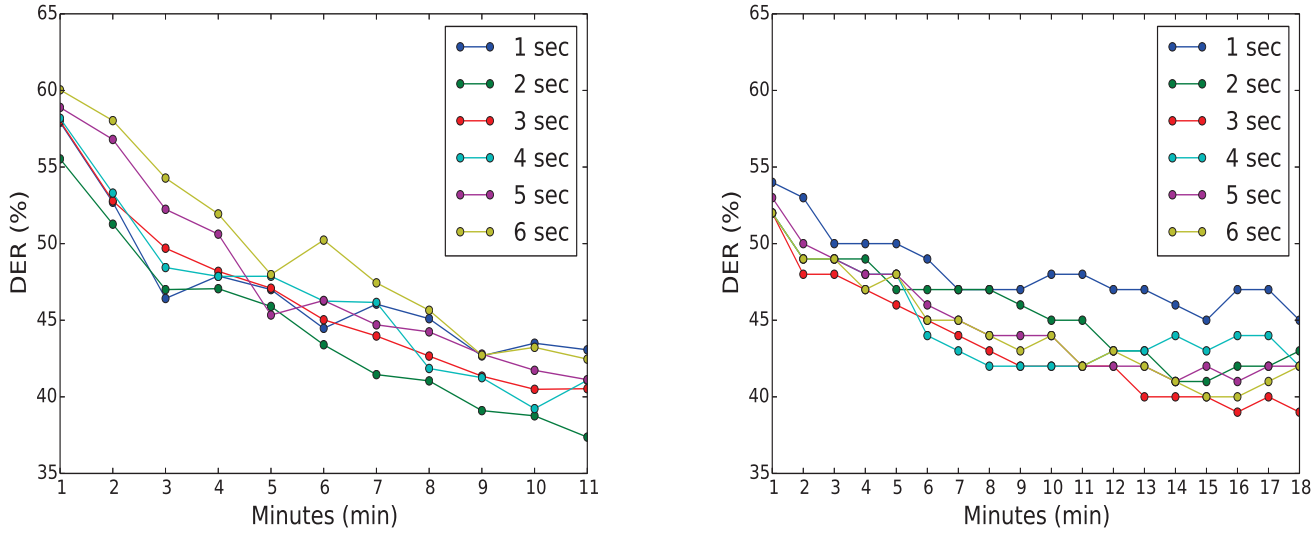

Fig. 3: An illustration of the evolution in DER against time (minutes) and for different segment durations. Results shown for the RTdev (left) and RTeval (right) datasets

5.4. Future work

These observations suggest that a speaker penalty might be useful to favour the introduction of fewer speakers, or at least reduce the rate at which new speakers are added. Of course, this is likely to reduce the overall convergence rate; the full number of speakers will take longer to appear in the hypothesis.

Two alternative strategies might help to reach a better compromise between the overestimation of speaker numbers and slower convergence. First, an adaptive speaker penalty may be used initially to favour the introduction of more speakers, but penalise their introduction later. Second, parallelised agglomerative hierarchical clustering may be applied periodically to merge similar models which might correspond to the same speaker. Since the computational implications of this approach may be too great for our intended application, their investigation has been left for future work.

Finally, since the use of longer segments introduces latency, it will also be of interest to combine online diarization with phone

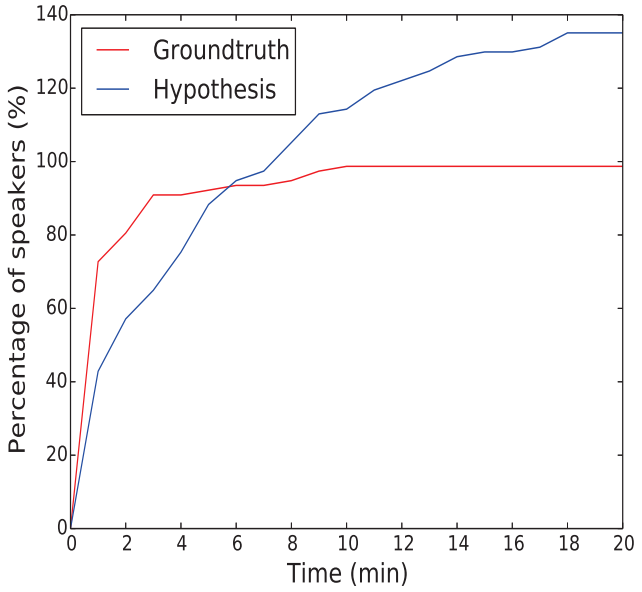

Fig. 4: An illustration of the evolution in speaker numbers for the RTdev dataset. Profiles shown for the ground-truth reference (red profile) and diarization hypothesis (blue profile).

adaptive training (PAT) [14, 15] which can improve speaker modelling when dealing with short utterances or segments. Since PAT involves only a linear feature transform, the additional computational requirements are minimal.

6. CONCLUSION

This paper presents a new adaptive, online approach to speaker diarization. In contrast to the past work, this paper addresses the most challenging of domains, namely meeting data characterised by high speaker overlap, spontaneity and short speaker turns. Experiments show that the best performance implies a latency in the order of 2 seconds and that speaker models convergence as the amount of training data increases. While results are in line with those reported for less challenging data, diarization error rates remain high. The paper also presents some strategies to improve performance through future work. This should focus not only on reducing diarization errors, but also on the rate of convergence, more reliable estimates of speaker numbers and improved diarization in the case of short segments.

7. REFERENCES

[1] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 5, pp. 1557–1565, Sept 2006.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[3] N. Evans, S. Bozonnet, Dong Wang, C. Fredouille, and R. Troncy, "A comparative study of bottom-up and top-down approaches to speaker diarization," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 382–392, 2012.

[4] S. Bozonnet, N.W.D. Evans, and C. Fredouille, "The LIA-Eurecom RT'09 speaker diarization system: enhancements in speaker modelling and cluster purification," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2010, pp. 4958–4961.

[5] G. Friedland, A. Janin, D. Imseng, X. Anguera, L. Gottlieb, M. Huijbregts, M.T. Knox, and O. Vinyals, "The ICSI RT'09 speaker diarization system," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 371–381, 2012.

[6] J. T. Geiger, F. Wallhoff, and G. Rigoll, "GMM-UBM based open-set online speaker diarization.," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2010, pp. 2330–2333.

[7] K. Markov and S. Nakamura, "Never-ending learning system for on-line speaker diarization," in *IEEE Workshop Automatic Speech Recognition Understanding (ASRU)*, Dec 2007, pp. 699–704.

[8] K. Markov and S. Nakamura, "Improved novelty detection for online gmm based speaker diarization," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2008, pp. 363–366.

[9] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 499–513, Feb 2012.

[10] S. Araki, T. Hori, M. Fujimoto, S. Watanabe, T. Yoshioka, T. Nakatani, and A. Nakamura, "Online meeting recognizer with multichannel speaker diarization," in *Conf. Signals, Systems and Computers (ASILOMAR)*, Nov 2010, pp. 1697–1701.

[11] D. Liu and F. Kubala, "Online speaker clustering," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, May 2004, vol. 1, pp. I – 333–6.

[12] C. Vaquero, O. Vinyals, and G. Friedland, "A hybrid approach to online speaker diarization," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2010, pp. 2638–2641.

[13] T. Oku, S. Sato, A. Kobayashi, S. Homma, and T. Imai, "Low-latency speaker diarization based on bayesian information criterion with multiple phoneme classes," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, March 2012, pp. 4189–4192.

[14] S. Bozonnet, R. Vipperla, and N. Evans, "Phone adaptive training for speaker diarization," in *Proc. Ann. Conf. International Speech Communication Association (INTERSPEECH)*, 2012.

[15] G. Soldi, S. Bozonnet, Alegre F., C. Beaugeant, and N. Evans, "Short-duration speaker modelling with phone adaptive training," *Odyssey*, 2014.