Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 1065–1103

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A review on speaker diarization systems and approaches

M.H. Moattar *, M.M. Homayounpour

Laboratory for Intelligent Multimedia Processing (IMP), Computer Engineering and Information Technology Department, Amirkabir University of Technology, Tehran, Iran

Received 12 December 2010; received in revised form 24 February 2012; accepted 29 May 2012
Available online 6 June 2012

Abstract

Speaker indexing or diarization is an important task in audio processing and retrieval. Speaker diarization is the process of labeling a speech signal with labels corresponding to the identity of speakers. This paper includes a comprehensive review on the evolution of the technology and different approaches in speaker indexing and tries to offer a fully detailed discussion on these approaches and their contributions. This paper reviews the most common features for speaker diarization in addition to the most important approaches for speech activity detection (SAD) in diarization frameworks. Two main tasks of speaker indexing are speaker segmentation and speaker clustering. This paper includes a separate review on the approaches proposed for these subtasks. However, speaker diarization systems which combine the two tasks in a unified framework are also introduced in this paper. Another discussion concerns the approaches for online speaker indexing which has fundamental differences with traditional offline approaches. Other parts of this paper include an introduction on the most common performance measures and evaluation datasets. To conclude this paper, a complete framework for speaker indexing is proposed, which is aimed to be domain independent and parameter free and applicable for both online and offline applications.

Keywords: Speaker indexing; Speaker diarization; Speaker segmentation; Speaker clustering; Speaker tracking

Contents

* Corresponding author. Tel.: +98 2164542722; fax: +98 2166495521.
E-mail addresses: moattar@aut.ac.ir (M.H. Moattar), homayoun@aut.ac.ir (M.M. Homayounpour).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.05.002

1. Introduction

Nowadays, there is a growing interest towards applying speech and language technologies in automatic searching, indexing, and retrieval of audio information. Automatic indexing and retrieval of information is possible via extracting meta-data from content. Several technologies are necessary for extracting meta-data. An audio document consists of multiple audio sources. Audio sources may be different speakers, music segments, types of noise, and etc. For instance, a broadcast news program consists of speech from different speakers as well as music segments and commercials. Audio indexing is the task of labeling audio segments within an audio document. The simplest audio indexing task concerns classifying the audio signal to speech and non-speech, in which non-speech is a general class consisting of music, silence, noise, etc. More complicated indexing would further mark the speaker changes in the detected speech and cluster segments of speech coming from the same speaker. This task is usually referred to as speaker indexing (Meignier et al., 2006) and is the focus of most current research efforts in audio indexing.

Speaker indexing is a process in which labels associating with speaker identities are assigned to different parts of an audio file. In speaker indexing, usually there is no primary model for speakers and the system should operate in an open-set manner. This means that the system should first determine if there exists a trained model of the current speaker in the model set or the utterance corresponds to a new speaker. Therefore, unsupervised speaker indexing may be required. Speaker indexing is an integral element of content based data mining applications.

Speaker indexing consists of two phases. In the first step, the audio document is segmented according to the speaker changes. This segmentation is performed in a manner that each speech segment contains the utterance of a single speaker. This stage is called speaker segmentation in the literature. In the second step, which is referred to as speaker clustering, the whole speech file is traversed, and all the speech segments uttered by the same speaker are labeled identically. Online speaker clustering, in which the speech segments are assigned a label as soon as the next change point is detected, is referred to as speaker tracking in literature. Speaker segmentation usually precedes speaker clustering. In such a case, the segmentation error degrades clustering performance. However, there are some approaches that optimize both speaker segmentation and clustering (Meignier et al., 2006; Meignier et al., 2001; Ajmera et al., 2002; Zhu et al., 2005; Barras et al., 2006). Speaker segmentation followed by speaker clustering is also called diarization (Meignier et al., 2006; Sinha et al., 2005; Tranter and Reynolds, 2006).

Diarization has received much attention recently, and specific competitions under the license of the National Institute of Standards and Technology (NIST) are devoted to it. Some applications of speaker diarization are:

- Movie analysis: movie analysis concerns different tasks. For example, dialogue detection determines whether a dialogue occurs in an audio recording or not. Also, questions such as who the speakers are or when actors appear, could also be addressed.
- Rich transcription: rich transcription (RT) (Meignier et al., 2006; Gales et al., 2006) adds several metadata in a spoken document, such as speaker identity and sentence boundaries. By indexing the audio according to the speakers and adding extra information to speech transcriptions, it becomes easier for humans to locate information and for machines to process.
- Automatic speech recognition (ASR) systems: segmentation algorithms can be used to split the audio into small segments for the ASR systems to process. Also, speaker diarization algorithms can be used to cluster the input data into speaker specific clusters for model adaptation and ASR performance improvement.
- Audio indexing and retrieval: a speaker diarization system allows automatic indexing of spoken audio documents, enabling the end user to browse the audio document by the identity of the speakers or their count.
- Audio archiving and monitoring: having archived the meetings or conferences, they could be easily accessed and monitored by interested individuals who were unable to attend such meetings.
- Speaker counting: this application involves determining the number of speakers participating in a conversation (most likely without having any a priori information about any of the speakers). Speaker counting can be employed in criminal activity detection; for example, in prisons, where three-sided call is prohibited, detecting the presence of a third speaker in recorded conversations could be helpful in identifying violators.
- Call routing: another application of speaker detection and tracking is automatic call routing based on the caller identity. For call routing, an incoming call could be identified as from a previous or new customer. Previous customers would be handled in a personalized manner, whereas new customers' information would be entered into the system and new speaker models are automatically generated.

According to Reynolds and Torres-Carrasquillo (2004) there are three main application domains for speaker diarization:

- Broadcast news (BN): radio and TV programs with various kinds of contents, usually containing commercial breaks and music, over a single channel.
- Meetings: meetings or lectures where multiple people interact in the same room. Normally recordings are made with several microphones. Meeting speech data differ by the number and location of the microphones. If only one microphone is used for recording the speech of all participants, the input format is single distant microphone (SDM). This type of meeting recordings is fundamentally similar to broadcast news. If there is more than one microphone in different locations of the meeting room, the recording condition is called multiple distant microphones (MDM). If every participant has his special microphone placed close to her/him, the recording condition is called individual personal microphone (IPM) or individual head microphones (IHM).
- Conversational telephone speech (CTS): single channel recordings of telephone conversations between two or more people.

The data from these domains differ in the quality of the recordings, the amount and types of non-speech sources, the number of speakers, the duration of speaker turns, and the style of the speech. Some of the main differences between indexing in these three different domains are mentioned in Table 1. Each domain presents unique challenges, although some techniques tend to generalize over several domains (Moh et al., 2003; Anguera et al., 2005).

The indexing task is also defined by the amount of prior knowledge allowed. There may be specific prior knowledge via sample speech from the speakers in audio signal and hence the task becomes similar to a speaker detection task (Martin and Przybocki, 2001). Prior knowledge could be example speech obtained from some of the speakers such as regular anchors on news, or knowledge of the number of speakers in the audio, perhaps for a teleconference over a known number of lines. Most of this prior knowledge can be used to improve diarization performance (Moraru, 2004). However, it is more desired to operate without any specific prior knowledge of the audio.

Most of diarization systems perform the task in a straight framework which contains some key components. The flow diagram of a conventional diarization system is presented in Fig. 1. A particular speaker diarization system starts with speech/non-speech detection or sometimes simply by just a silence removal. In MDM meeting conditions, silence removal may be preceded by acoustic beamforming; however, the latter task is not general. These two tasks can be summarized under a single preprocessing operation. The next operation is feature extraction which is a key part of all speech and signal processing systems. The following step is to segment the speech stream by the speaker changes. This task may be succeeded by a change point refinement or re-segmentation step which enhances the change point locations or reduces false alarms. After segmentation or in parallel with this phase, speech segments will be clustered according to the speaker characteristics. Speaker-based clustering can also be followed by cluster re-combination or re-organization, which refines the speaker-based clusters for more homogeneity. The result of this step can also have a positive impact on change point

Table 1
Differences between speaker indexing for meetings, broadcast news and telephony conversations.

Telephony conversations	Broadcast news	Meeting speech
Number of speakers limited to 2 or at most 3 persons	Number of speakers could be 10 persons or more	The number of speakers is limited to the capacity of the meeting room
Music and other audio contents do not exist	Some parts of the file may contain music or commercials	music or commercials does not exist
Recording channel and environment do not usually change	The recording condition of each speaker may vary	Variations in recording quality, including impulse noises, reverberation and variable speech levels may exist
The recording channel and environment are different for each speaker	The recording channel and environment may be different for each speaker	All the conversations take place in one place
Average speaker change duration is usually so short	Average speaker change duration is longer	Average speaker change duration may be short
Normal existence of overlapping regions where two or more speakers speak simultaneously	Normally there is no overlapping regions between speaker utterances	Normally there are overlapping regions between the speech of two speakers
Only one recording microphones is applied	the recording is performed using several microphones	The recordings may be performed with one or two channels

false detections. Usually speaker diarization terminates by speaker cluster formation, but may be followed by speaker-based speech labeling and indexing.

Speaker diarization is an appealing research area and numerous groups and centers have been investing on it (Kotti et al., 2008). As an example, the Global Autonomous Language Exploitation (GALE) program (International Computer Science Institute, xxxx), which is contributed by the speech group at the International Computer Science Institute at Berkeley, concerns speaker diarization, sentence segmentation, machine translation, and information distillation in various languages. A well-known center which is active in developing speaker diarization systems is International Computer Science Institute (ICSI, Berkeley, California) whose efforts have resulted in the ICSI-SRI diarization system (Anguera et al., 2005). This system is implemented in different versions and has participated in some of NIST rich transcription evaluations including RT06 and RT07. This system is based on agglomerative clustering and automatically deduces the number of speakers in a recording, along with the information about where each speaker is speaking. The Laboratoire d'Informatique d'Avignon (LIA) system and the Communication Langagiere et Interaction Personne Systeme-Institut d'Informatique et Mathematiques Applique'es de Grenoble (CLIPS) system (Moraru, 2004) are other examples for the systems developed for speaker segmentation and clustering. Both LIA and CLIPS labs are members of the ELISA consortium, and have participated in NIST rich transcription evaluations since 2000 (Moraru, 2004; Moraru et al., 2003; Moraru et al., 2004). Speaker segmentation and location-based multichannel speaker segmentation has also been widely investigated by the Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP) (IDIAP Research Institute, xxxx). Augmented multi-party interaction (AMI) (AMI, xxxx) is one of the projects undertaken by IDIAP, which is concerned with real-time human interaction in smart meeting rooms. Two other widely referred diarization projects are CUED system developed in Cambridge University Engineering Department (Reynolds and Torres-Carrasquillo, 2004) and MIT-LL system implemented in the Massachusetts Institute of Technology-Lincoln Labs (Sinha et al., 2005) which perform automatic segmentation, clustering and labeling of speakers on broadcast news data.

The Transonic solutions project, from the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (SAIL, xxxx), also deals with speech segmentation. The Spoken Language Processing Group in the Computer Sciences Laboratory for Mechanics and Engineering Science (LIMSI-CNRS) at Paris has also invested effort on rich transcription of multilingual spoken documents (LIMSI, xxxx). The Department of Speech, Music and Hearing of the Royal Institute of Technology (KTH) at Stockholm is also dedicated to speaker segmentation as a preprocessing step of the human–computer interaction task (Department of Speech and and Hearing, xxxx). A similar case is Computers in the Human Interaction Loop (CHIL) project (CHIL, xxxx). Microsoft

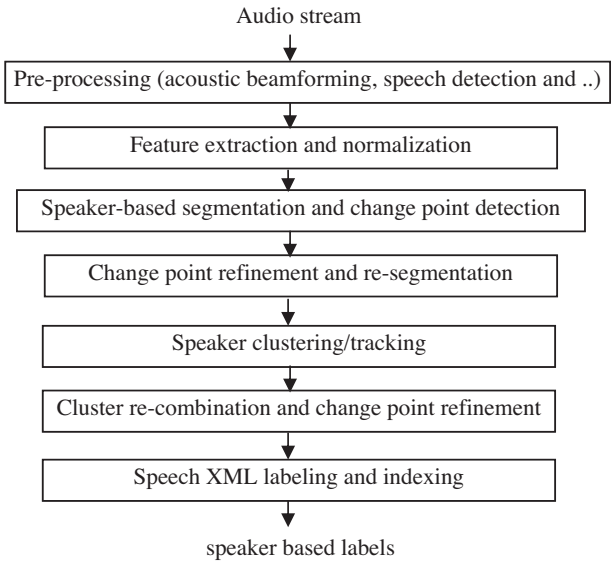

Fig. 1. A general speaker diarization framework.

research has also been active in speaker segmentation (Microsoft Audio Projects, xxxx), as a part of the Audio Content Analysis project, where discrimination among six audio classes including pure speech, speech over music, speech over noise, pure music, background sound and pause/silence is considered. Finally, I6-Aachen group has developed an automatic segmentation algorithm for MPEG audio streams which targets at developing a toolbox for content analysis and video retrieval (The Chair of Computer Science, xxxx).

Also, several toolkits distributed under open-source licenses are available on the web. One of the oldest is the CMU segmentation toolkit which was released in 1997 (Siegler et al., 1997) and was developed for the NIST broadcast news evaluation campaign. AudioSeg (Gravier et al., 2010), under the GPL license, is a toolkit developed by IRISA during the ESTER 1 campaign in 2005. It includes audio activity detector, segmenting and clustering tools and a Viterbi decoder (Gravier et al., 2010). Mistral (xxxx), under the GPL license, is also dedicated to speaker diarization and relies on the ALIZE speaker recognition library. Another open-source diarization library is LIUM SpkDiarization (Meignier and Merlin, 2010) toolkit which is published under the GPL license. LIUM SpkDiarization is oriented towards speaker diarization of broadcast news. The toolkit provides elementary tools, such as segment and cluster generators and decoder and model trainers. Fitting those elementary tools together is an easy way of developing a specific diarization system, but these components can also be employed for other tasks. Different modules are implemented in this toolkit which include, feature extraction, speech detection, gender and bandwidth detection, speaker segmentation, speaker clustering and Viterbi based speaker segmentation.

This paper offers the state-of-the-art approaches and algorithms for speaker indexing and diarization. The main goal of this paper is to summarize the most widely proposed approaches and the biggest milestones, besides offering a complete overview on the most important aspects of speaker indexing and diarization. This paper is organized as follows: Section 2 introduces the databases commonly used in diarization performance evaluation. Conventional acoustic beamforming approaches for meeting conversations is explained in Section 3. In Section 4, the feature extraction problem and the most commonly used features for speaker indexing are introduced. Section 5 discusses the speech activity detection as one of the most important diarization front-end processing. Section 6 presents a review on speaker segmentation approaches and achievements. A detailed description of the state-of-the-art clustering approaches is offered in Section 7. Section 8 discusses the online indexing and the algorithms proposed to tackle this problem. Section 9 reviews the systems and approaches which consider the indexing task as a whole and concern both speaker segmentation and speaker clustering in a unified framework. Section 10 introduces the area of audio-visual diarization and indexing. The main indexing performance measure as introduced by NIST rich transcription (RT) evaluations is mentioned in Section 11. Our proposed unified speaker indexing framework is introduced in Section 12. Section 13 introduces the directions for future works. Finally, Section 14 discusses the conclusions.

2. Evaluation databases

Since the characteristics of diarization application domains are different, any application specific approach needs to be evaluated on a specified database. Therefore, speaker diarization evaluation databases are divided to application specific types. Apart from locally recorded or artificially created databases, the most common evaluation datasets are introduced in this section.

2.1. Broadcast news database

Hub-4 is one of the most widely used speech corpora for evaluating broadcast news (BN) diarization systems. NIST 1996 HUB-4 evaluation dataset consists of 4 files of almost half an hour (Stern, 1997). Hub-4 1997 English Broadcast News Speech Database is composed of about 97 hours of news broadcasting from different radio stations such as CNN, ABC, CRI, and C-SPAN. 1997 Mandarin Broadcast News Speech Corpus (Hub4-NE) contains recorded broadcasts from CCTV, KAZN and VOA. The main content of this data are news reports and interviews. There are short 1 sec segments as well as long reports of more than 30 sec in the conversations. Each file is approximately 30 min long, and contains more than 10 speaker's utterances. The environment in this database is close to real life. There are background noises, short time noises, and even background occasional music in the conversations.

Also, 1998 HUB-4 broadcast news evaluation english test material (Alabiso et al., 1997) consists of television and broadcast news speech recorded at 16 KHz sampling rate. The evaluation test data consist of approximately three hours of speech divided into two datasets. The first dataset (set1) is taken from broadcasts between October 15 and November 14, 1996. The second set (set2) is taken from different variety of shows broadcast in June, 1998. Each of these datasets is selected by NIST from larger pools collected by the LDC. The data consist of a single monophonic channel, and commercials or sports are excluded from the evaluations; since they contain linguistic and acoustic challenges which are considered to be outside of the evaluation domain.

ESTER SD benchmark (Galliano et al., 2006) consists of 32 shows from various France Radio Channels. The acoustic excerpts come from six different sources. The first part of this audio corpus is made up of a total of 100 hours of radio broadcast news shows recorded in 1998, 2000, 2003 and 2004, which are manually transcribed. The second part is made up of 1677 hours of non transcribed shows recorded between October 2003 and September 2004. Audio files are recorded in wave format at 16 KHz

16-bits from a standard audio card on a PC without any compression. 90 hours of the transcribed data are devoted to training and development while the 10 remaining hours compose the test set. The non transcribed data is also made available for training, in order to investigate the improvements that can be achieved by using large amount of such data. Another widely used speech database for broadcast news evaluation is GALE Mandarin dataset (Chu, 2008) which contains 1900 hours of broadcasting news speech collected from various TV programs. The waveforms are sampled at 16 KHz and quantized at 16 bits per sample.

2.2. Meeting speech database

In meeting domain, the most widely used corpus is ICSI Meetings Corpus (Janin et al., 2003): which contains 75 meetings with about 72 hours in total. They are recorded in a single meeting room, with 4 omni-directional tabletop and 2 electret microphones. Also, CMU Meeting Corpus (Burger et al., 2002) includes 104 meeting conversations of an average duration of 60 min with 6.4 participants per meeting. These conversations are focused on a given scenario or topic. Initial meetings have 1 omni-directional microphone and newer ones have 3 omni-directional tabletop microphones.

NIST Pilot meeting corpus (NIST Pilot Meeting Corpus, 2006) is another database dedicated to meeting speech which consists of 19 meetings with a total of 15 hours. Audio recordings are done using 3 omni-directional table-top microphones and one circular directional microphone with 4 elements. Another example is CHIL meeting corpus in which recordings were conducted in 4 different meeting room locations. Each meeting room is composed of several distant microphones, as well as speaker localization microphones and microphone arrays. Each meeting also contains several video cameras. Also, AMI meeting speech corpus (AMI, xxxx) includes 100 hours of meetings with generally 4 participants. These are split into two main groups: real meetings and scenario-based meetings (where people are briefed to talk about a particular topic). In the corpus, one or more circular arrays of 8 microphones are centrally located on the table.

2.3. NIST rich transcription evaluation

More general evaluation corpora are released by NIST (NIST Rich Transcription evaluations, 2006; NIST Fall Rich Transcription, 2006) as the benchmarks for the NIST rich transcription evaluations. NIST has been organizing multiple evaluations on many aspects of speech technologies over the years. In the area of speaker diarization evaluations, they started in year 2000 with interest in telephony speech (2000, 2001, and 2002), then broadcast news (2002, 2003, and 2004) and more recently meetings (2002, 2004, 2005, and 2006). In the recent years, their focus has been turned towards the meeting environment. A common characteristic of these evaluations is that the only a priori knowledge available to the participants is the knowledge about the recording scenario/source (e.g. conference meetings, lectures, or coffee breaks for the meetings domain), the language (English), and the formats of the input and output files. Evaluation participants may use external or background data for building world models and/or for normalization purposes but no a priori information relating to speakers in the recordings is available (NIST Rich Transcription evaluations, 2006).

Initially, in 2002, the speaker segmentation evaluation was held within the speaker recognition evaluation (SRE-02). This routine was changed in 2004 to 2006, when speaker diarization has been a part of the rich transcription (RT) evaluation (RT04s, RT05s and RT06s), grouped with speech-to-text evaluation (STT). The datasets used for these evaluations contain data from CMU, ICSI, LDC, NIST, CHIL, VT, EDI, IDI, TNO and AMI. Two types of meetings are recorded: lecture and conference (NIST Rich Transcription evaluations, 2006).

Some example NIST RT evaluation databases are as follows: NIST 2002 Rich Transcription Broadcast News (BN) and Conversational Telephone Speech (CTS) (Garofolo et al., 2002) (NIST RT'02) consists of the speech files for evaluating rich transcription systems in the domains of broadcast news and conversational telephone speech. The CTS data is composed of 60 approximately five-minute excerpts from 60 different conversations: 20 excerpts from Switchboard-1 data, 20 excerpts from Switchboard-2 data, and 20 excerpts from Switchboard Cellular-2 data. The BN data is composed of six approximately 10-min excerpts from six different broadcasts. The broadcasts are selected from programs from MNB, PRI, NBC, CNN, VOA and ABC, all collected in 1998. NIST RT'04 speaker diarization data (Fiscus et al., 2004) consists of one 30-min extract from 12 different U.S. broadcast news shows. These speech files were derived from TV shows: three from ABC, three from CNN, two from CNBC, two from PBS, one from CSPAN, and one from WBN. The style of show varies from a set of lectures from a few speakers (CSPAN) to rapid headline news reporting.

In recent years, the NIST RT evaluations have focused on the conference meeting domain, where the spontaneous speaking style presents a considerable challenge for speaker diarization. Each meeting used in the evaluations was recorded using multiple microphones which are positioned on the participants or in different locations around the meeting room. By grouping these microphones into different classes, NIST created several evaluation conditions. These include: individual headphone microphones (IHM), single distant microphones (SDM), multiple distant microphones (MDM) and all distant microphones (ADM).

Participating teams are required to submit a hypothesis of speaker activity including start and end times of speech segments with speaker labels, but do not need to reflect the speaker's real identity. The outputs are compared with the ground-truth reference in order to obtain the overall detection error rate (DER) as discussed in Section 11. The DER

metric is the sum of three sources of error: missed speech (percentage of speech in the ground-truth but not in the hypothesis), false alarm speech (percentage of speech in the hypothesis but not in the ground-truth) and speaker error (percentage of speech assigned to the wrong speaker). The speaker error can be further classified into incorrectly assigned speakers and speaker overlap error. In the first case the hypothesized speaker does not correspond to the real (ground-truth) speaker. Speaker overlap error refers to the case when wrong number of speakers is hypothesized when multiple speakers speak at the same time. When comparing the system outputs with the reference, the scoring algorithm computes an optimum mapping between both sets of labels in order to obtain the DER.

References for evaluating speaker diarization are initially obtained via manual labeling of the acoustic data; however, high variations between different labelers can be problematic. Therefore, more recently, an automatically generated forced alignment has been used in order to extract more reliable speaker start and end points using automatic speech recognition (ASR).

The latest version of the NIST rich transcription evaluations is NIST RT'09 (Rich Transcription Meeting Recognition Evaluation Plan, 2009). Like the other recent evaluations, this evaluation only includes meeting conversations. The term of participation and the evaluation rules and metrics are the same as the previous version and the only difference is the amount of test data which is used in evaluations. Also, in this evaluation no lecture test set is considered and the speech signals are associated with their corresponding video for evaluating multi-model systems. The training and development corpora for this evaluation include ICSI Meeting Corpus, ISL Meeting Corpus, NIST Meeting Pilot Corpus, NIST Phase II Meeting Corpus, CHIL '05, '06, and '07 development test sets, AMI Meeting Corpora, RT-04S Development and Evaluation Data, RT05, RT06 and RT07 Evaluation Data and Fisher English conversational telephone speech corpus.

There are differences between the sets of meetings used each year. For RT'05 the average speaker segment duration is 2.5 sec. This value decreases continuously for subsequent datasets (2.3 sec for RT'06, 2.0 sec for RT'07 and 1.8 sec for RT'09). This tendency leads to increasingly more frequent speaker turns and increases the chances of miss-classifying a speech segment. The average segment duration for RT'05 is 2.1 sec. This value falls to 1.4 sec for RT'06, RT'07 and RT'09. The consistent decrease in speaker/turn duration explains the differences in results from one dataset to another.

There are also noticeable differences in silence and overlap statistics. The percentage of silence is lower for RT'05 and RT'09 datasets than RT'06 and RT'09 datasets. However, the RT'05 and RT'09 datasets have a higher overlap rate than RT'06 and RT'07 datasets. Average overlap in RT'09 dataset is slightly higher compared to RT'05. Conversely, RT'05 and RT'09 have lower average percentage of silence compared to RT'06 and RT'07. Lower silence rate and higher overlap indicate that these meetings are more dynamic, with less idle time and more discussion.

Comparisons on the works of the speech and speaker recognition communities highlight the need for huge datasets. It is apparent that the lack of large speaker diarization datasets makes it difficult to assess novel algorithms. Significantly larger datasets are needed in order to obtain more robust and meaningful performance estimates and comparisons.

3. Acoustic beamforming

Applications of speaker diarization to the meeting domain triggered the need for dealing with multiple microphones which are often used to record the same meeting from different locations in the room (19, 24, and 35). The microphones can have different characteristics: wall-mounted microphones (intended for speaker localization), lapel microphones, desktop microphones positioned on the meeting room table or microphone arrays. The use of different microphone combinations as well as differences in microphone quality made it necessary to investigate new approaches for speaker diarization with multiple channels.

A variety of algorithms have been proposed to extend mono-channel diarization systems to handle multiple channels. One option, proposed in (Fredouille et al., 2004), is to perform speaker diarization on each channel independently and then to merge the individual outputs. In order to do so, a two axis merging algorithm is used which considers the longest detected speaker segments in each channel and iterates over the segmentation output. A late-stage fusion approach was also proposed in (Jin et al., 2004), in which speaker segmentation is performed separately in all channels and diarization is applied with only taking into account the channel whose speech segments have the best signal-to-noise ratio (SNR). Subsequent approaches investigated pre-processing to combine the acoustic signals to obtain a single channel which could then be processed by a regular mono-channel diarization system. In (Istrate et al., 2005) multiple channels are combined with a simple weighted sum according to their SNR.

Since the NIST RT'05 evaluation, the most common approach for multi-channel speaker diarization involves acoustic beamforming as proposed in (Anguera et al., 2005) and described in (Anguera et al., 2006). There are two main groups of beamforming techniques that can be found in the literature. These type are called data-independent (or fixed) and data-dependent (or adaptive). The techniques that are data-independent fix their parameters and maintain them throughout the processing of the input signal. Data dependent techniques update their parameters to better suit the input signal, adapting to changing noise conditions. Moreover, there are several post-processing techniques that are applied after the beamforming, some of them are very linked to the beamforming process.

Fixed beamforming techniques are simpler to implement than the adaptive ones, but are more limited in their ability to eliminate noise sources. The simplest beamforming technique in this group is the delay & Sum (D&S) technique (Flanagan, 1994; Johnson and Dudgeon, 1993). The D&S beamforming is a particular case of a more general definition of a filter & sum beamforming where an independent filter is applied to each channel. One application of such techniques is super directive beamforming (SDB) (Cox et al., 1986; Cox et al., 1987), where some channel filters are defined to maximize the array gain, which is defined as the improvement in SNR between the reference channel and the "enhanced" system output. For the case of near-field signals (like when a microphone array is located right in front of the speakers) the SDB is reformulated by using near-field propagation functions for acoustic waves (McCowan et al., 2000). Considering the speech signal to be narrow-band simplifies the design of beamforming systems but does not represent the reality well. To deal with broadband signals in an effective manner, several sub-array beamforming techniques have been proposed (Sanchez-Bote et al., 2003; Fischer and Kammeyer, 1997) in which the set of microphones is split into several sub-arrays which focus their processing in a particular band, collapsing all the information into the "enhanced" signal at the end.

The adaptive beamforming techniques present higher capacity at reducing noise but are much more sensitive to errors due to the approximation of the channel delays. The Generalized Sidelobe Canceller (GSC) technique (Griffiths and Jim, 1982) aims at enhancing the signal that comes from the desired direction while cancelling out signals coming from other sources. This is achieved by creating a double path for the signal in the algorithm. A standard beamforming path consists of a blocking matrix and a set of adaptive filters that aim at minimizing the output noise power. The blocking matrix blocks the desired signal from the second path. At the end, both paths are subtracted to obtain the output signal. In order to find the optimum coefficients for the lower part, an algorithm like the least mean squares (LMS) can be used. Although widely used, in practice the GSC can suffer from distortion. This is due to the inability of the blocking matrix to completely eliminate the desired signal from the adaptive path. This problem is treated in (Hoshuyama et al., 1999) where the blocking matrix is designed with control of the allowed target error region. These approaches have high computational requirements and there is the risk of converging to inaccurate parameters, especially when processing microphones of different types.

Other kinds of adaptive beamforming techniques are those that allow a small amount of distortion of the desired signal. One of these techniques is named the AMNOR (Adaptive microphonearray system for noise reduction), introduced by Kaneda (1991), Kataoka and Ichirose (1990). It introduces a known signal during noise-only periods in order to adapt the filters to cancel such signal and therefore improve the quality of the speech parts. One drawback of this technique is the need for accurate speech/non-speech detection. Some efforts have been reported that apply adaptive beamforming techniques to the nearfield case. In (McCowan et al., 2000), adaptive beamforming and super-directive beamforming techniques are combined for this purpose.

In real applications none of the above beamforming techniques achieves the levels of improvement on the signal. In practice a post-processing of acoustic signal is necessary in order to obtain the optimum output quality. In (Zelinski, 1988) a Wiener post-filtering is applied where time delay information is used to further enhance the signal in the filter. In (Marro et al., 1998), it does an analysis of the interaction of Wiener filtering with a filter & sum beamforming, showing that the post-filter can cancel noise and allows slight errors. Other post-filtering approaches applied to microphone arrays beamforming are proposed in (Cohen and Berdugo, 2002, Valin et al., 2004). There are many post-processing techniques aimed to the single channel signal obtained from the beamforming process. Some of these techniques take into account acoustic considerations (Rosca et al., 2003; Zhang et al., 2004) or acoustic models (Brandstein and Griebel, 2001) to enhance the signal better.

Many RT participants use the free and open-source acoustic beamforming toolkit known as BeamformIt (Anguera, xxxx) which consists of an enhanced delay-and-sum algorithm. Speech data can be optionally preprocessed using Wiener filtering to remove noise. A reference channel is selected and the other channels are appropriately aligned and combined with a standard delay-and-sum algorithm. The contribution made by each signal channel to the output is then dynamically weighted according to its SNR or by using a cross-correlation-based metric. Various additional algorithms are available in the BeamformIt toolkit to select the optimum reference channel. Note that, although there are other algorithms that can provide better beamforming results, delay-and-sum beamforming is the most reliable one when no information on the location or nature of each microphone is known a priori.

4. Feature extraction

Feature extraction is one of the main parts of a signal processing application and can cause a significant impact on the system performance. Features extracted from the acoustic signal are intended to convey information about the speakers in the conversations in order to enable the systems to separate them optimally. According to Kinnunen and Li (2010), appropriate features for speaker modeling and discrimination should have the following properties:

- Have large between-speaker variability and small within-speaker variability.
- Be robust against noise and distortion.
- Occur frequently and naturally in speech.
- Be easy to measure from speech signal.

- The number of features should be relatively low.

Previous works have exploited different features for the speaker indexing task (Lu and Zhang, 2002; Lu and Zhang, 2005). Mel-frequency cepstral coefficients (MFCCs), sometimes with their first and/or second derivatives are the most common features (Sinha et al., 2005; Lu and Zhang, 2002). However, there is not a compromise on the order of MFCCs. Tritschler and Gopinath (1999), Sian Cheng and min Wang (2003), Ajmera et al. (2004), and Cheng and Wang (2004) suggest 24-order MFCCs while (Kim et al., 2005) utilizes 23-order MFCCs. Also, 13-order MFCCs along with their first-order derivatives are utilized in (Kotti et al., 2006; Kotti et al., 2006), while Wu and Hsieh (2006) employs 12-order MFCCs along with their first-order derivatives. In (Wu and Hsieh, 2006), several MFCC orders are investigated before the 12-order MFCCs along with their first derivatives are chosen. In (Kotti et al., 2008), an effort is made to discover an MFCC subset that is more suitable to detect a speaker change. Also, there is no consensus with respect to first-order MFCC derivatives. While, first-order MFCC derivatives are claimed to deteriorate efficiency in (Delacourt and Wellekens, 2000), the use of first-order MFCC derivatives is found to improve performance in (Wu and Hsieh, 2006).

Other frequently applied features are: short-time energy (STE) (Meignier et al., 2006), zero-crossing rate (ZCR) (Lu and Zhang, 2002), pitch (Lu and Zhang, 2002; Lu and Zhang, 2005), spectrum magnitude (Boehm and Pernkopf, 2009), Line spectrum pairs (LSPs) (Lu and Zhang, 2002; Lu and Zhang, 2005) and perceptual linear prediction (PLP) cepstral coefficients (Tranter et al., 2004; Chu et al., 2009). Features based on phoneme duration, speech rate, silence detection, and prosody are also investigated in literature (Wang et al., 2003).

A common approach in speaker diarization is to combine diverse features in parallel and benefit their discrimination power. In general, fusion techniques increase the reliability and robustness of a system (Zhu and Rong, 2003). For example, in (Lu and Zhang, 2002; Lu and Zhang, 2005), speaker segmentation using MFCCs, LSPs, and pitch features fused in a parallel Bayesian Network is proposed. Yamaguchi et al. (2005) proposes a speaker segmentation system using energy, pitch frequency, peak-frequency centroid and peak-frequency bandwidth, and adds three new features including temporal stability of the power spectra, spectral shape and white noise similarities. All these three features are related to the cross correlation of the power spectrum of the signal.

Although the aforementioned parameterization techniques yield a good performance in speaker diarization systems, they do not usually focus on representing the information relevant to distinguishing between speakers and do not isolate such information from other interfering sources.

As suggested in (Shriberg, 2007), using long-term features can reveal individual characteristics of the speakers' voices as well as their speaking behavior, which cannot be captured by frame-based short-term cepstral analysis. Friedland et al. (2009) and Antolin et al. (2007) show that the combination of traditional short-term features (i.e. MFCCs) with prosodic and other long-term features can improve the diarizaton results. Therefore, it studies the speaker discriminability of 70 different long-term features and then, combines the best long-term features with short-term features to increase the accuracy of speaker diarization. The initial candidate features include pitch, energy, formants, harmonics-to-noise ratio (HNR), and long-term average spectrum.

On the other hand, different feature normalization approaches can be applied in speaker indexing, including RASTA-filtered cepstra mean and variance normalization (Reynolds and Torres-Carrasquillo, 2004) and cepstral mean normalization (CMN) (Zamalloa et al., 2010). In order to avoid the influence of background noises and other non-speaker related events, feature warping is proposed to change the shape of the probability density function (pdf) of the features to a Gaussian shape prior to their modeling (Pelecanos and Sridharan, 2001; Ouellet et al., 2005). These approaches have been applied successfully in (Sinha et al., 2005 and Zhu et al., 2006) for speaker diarization in broadcast news and meetings, respectively. Feature warping is found to be more effective than other normalization techniques in speaker verification task (Barras and Gauvain, 2003). It is claimed that feature normalization is necessary to obtain significant performance in speaker indexing (Sinha et al., 2005).

In MDM meeting diarization redundant information is available in comparison with SDM meeting diarization or BN speech data. However, sometimes all speech signals are combined into one (Zamalloa et al., 2010), from which some acoustic features are extracted. A source of information used in MDM scenarios is the information related to speaker localization (Ellis and Liu, 2004), such as the time delay of arrival (TDOA) features (Pardo et al., 2006). TDOA features permit short-term speaker segmentation but do not provide any speaker identity information. On the other hand, acoustic features provide long-term speaker identity but require minimum durations to build reliable acoustic models. The use of the TDOA between the microphones for speaker diarization has been used either independently (Ellis and Liu, 2004; Lathoud et al., 2004) or in combination with acoustics (Ajmera et al., 2004; Pardo et al., 2006). Regardless of the method used for the combination of these two feature streams, a weighting between them needs to be applied to take clustering and segmentation decisions.

Authors of Anguera et al. (2006) proposed a system that obtains the TDOA values by applying an acoustic beamforming to all available channels and then combines it with the acoustic features by a weighted sum at the log likelihood level. The weights needed to be tuned by hand using development data and were fixed for all meetings. This imposes a restriction on the number of different features

to use, as the search space grows exponentially with the number of applied streams. In (Pardo et al., 2006) it was first demonstrated that TDOA between channels could be mixed with spectral features to obtain improved performance over a base system that used only spectral features. This TDOA information combined with the MFCC information has been used by all systems in the latest Rich transcription evaluation (Rich Transcription Evaluation Project, 2002-2009). The shortcomings of TDOA methods are due to using distant microphones. There are noises and reverberations in the recordings and the results are not free from errors. In (Evans et al., 2009) a method was presented to improve inaccurate estimates of delays and increase speaker separation in delay-space.

5. Speech activity detection

Speech activity detection (SAD) or voice activity detection (VAD) is one of the most crucial frontend processing in a diarization framework. The output of other subtasks highly depends on the precision of this task. The aim of SAD is to find the regions of speech in the audio stream. Depending on the application domain, non-speech regions may consist of many acoustic classes such as silence, music, room noise, background noise, or cross-talk.

Silence can be removed using a phone recognizer (Sinha et al., 2005) or energy constraint (Tranter et al., 2004), or in a final stage processing using a word recognizer (Zhu et al., 2005). Regions which contain commercials, and thus are of no interest for the final output, can also be automatically detected and removed (Tranter et al., 2004; Johnson and Woodland, 2000). When the speech detection phase is run early in a system, or the output is required for further processing, it is more important to minimize speech miss than false alarm rates, since missed speech is unrecoverable in most systems.

The general approach used in SAD is maximum-likelihood classification with Gaussian mixture models (GMMs) trained on labeled training data (Tranter and Reynolds, 2006). The simplest system uses just speech/non-speech models such as Wooters et al. (2004), while Nguyen et al. (2002) uses four speech models for the possible gender/bandwidth combinations. Similar approach is proposed in (Sun et al., 2010) in which using two initial models, all frames are classified into speech and non-speech. The classified frames are then used to iteratively re-train the speech and non-speech GMMs based on the maximum a posteriori (MAP) approach. Noise and music are explicitly modeled in (Zhu et al., 2005; Reynolds and Torres-Carrasquillo, 2004; Gauvain et al., 1998) which have classes for speech, music, noise, speech music, and speech noise, while (Hain et al., 1998 and Sinha et al., 2005) use wideband speech, narrowband speech, music and speech music models.

The classes can also be broken down further, as in (Liu and Kubala, 1999), which has eight models in total, five for non-speech (music, laughter, breath, lip-smack, and silence) and three for speech (vowels and nasals, fricatives, and obstruents). When operating on unsegmented audio, Viterbi segmentation using the models is employed to identify speech regions. If an initial segmentation is already available, each segment is individually classified. Minimum length constraints (Wooters et al., 2004; Liu and Kubala, 1999) and heuristic smoothing rules (Reynolds and Torres-Carrasquillo, 2004; Nguyen et al., 2002) may also be applied.

Unsupervised methods for voice activity detection use robust features like the 4Hz modulation energy for speech detection. However, the drawback of using energy is that it is not possible to use this approach when the audio contains fragments with high energy levels that are non-speech. In (Kristjansson et al., 2005) some well known robust features are proposed for SAD which are based on autocorrelation of the signal or the characteristics of spectrum.

The main drawback of model-based approaches is their reliance on external data for the training of speech and non-speech models which makes them less robust to changes in acoustic conditions. Hybrid approaches have been proposed as a potential solution to achieve more reliable and context independent performance. In most cases, an energy-based detection is first applied in order to label a limited amount of speech and non-speech data for which there is high confidence in the classification. In a second step, the labeled data are used to train meeting-specific speech and non-speech models, which are subsequently used in a model-based detector to obtain the final speech/non-speech segmentation.

In (Anguera etal., 2006; Anguera et al., 2006), the authors first use a derivative filter in combination with a finite state machine (FSM) to detect speech and non-speech regions. These initial labels are then used to build two HMMs for speech/non-speech. The system iteratively segments and trains both models until the overall likelihood stops increasing. A more complex system is described in (Huijbregts et al., 2007; Wooters and Huijbregts, 2008) which is able to detect audible non-speech. El-Khoury et al. (2009); Nwe et al. (2010) uses the fusion of supervised GMM based techniques and unsupervised methods based on energy.

VAD approaches in telephone or meeting domain would be different from broadcast news application. For telephony audio, typically some form of standard energy/spectrum-based speech activity detection is used, although the GMM approach has also been successful in this domain with single channel (Tranter et al., 2004) or cross channel (Liu and Kubala, 2003) classes. For meeting audio, the non-speech can be from a variety of noise sources, like paper shuffling, coughing, laughing, etc. Also, energy based methods do not currently work well for distant microphones (van Leeuwan, 2005; Istrate et al., 2005), so using a simple pre-trained speech/non-speech GMM is preferred (Anguera et al., 2005; van Leeuwan, 2005; Istrate et al., 2005).

The problem of speech activity detection is so important in the meetings domain that a separate evaluation for speech activity detection was introduced in the spring 2005 Rich Transcription meeting evaluation (Fiscus et al., 2005).

5.1. Well-known SAD subsystems

There are also many well-known and widely referred SAD subsystems in literature. The SAD module of ICSI-SRI diarization system (Anguera et al., 2007; Anguera et al., 2006) is one of these VAD approaches. ICSI-SRI diarization system is available in two versions. In the 2005 version, speech/non-speech segmentation was performed using a model-based approach (Anguera et al., 2007). One model for speech and one model for non-speech were trained on a training set of meetings. Each model contained three states and the first twelve MFCC coefficients were used as input. The drawback of this system was that new models needed to be trained as soon as the conditions change. Therefore in 2006, the system was replaced by a two step algorithm (Anguera et al., 2006). First, a silence based set-up was used to find all segments with low energy. It was assumed that silence was the only form of non-speech in the meetings and that this first step was able to find enough representative speech and silence segments to be used in the second step. In the second step, the segments were used to train a model-based system with two states: one for speech and one for non-speech. Hidden Markov model (HMM) was used to realign the data. Using this new alignment, new GMMs were trained. After a number of iterations the final speech/non-speech alignment was obtained. The advantage of the 2006 system is that no training data is needed for the speech and non-speech models. The restriction is that this system is only able to distinguish between speech and silence. Because in the first step an energy-based segmentation is done, all non-speech with high energy will be classified as speech. In the second step this error will not be corrected. Fortunately, the assumption that all non-speech in the meetings under evaluation is silence is often valid. The structure of ICSI-SRI VAD system is illustrated in Fig. 2.

A similar approach is applied by LIA meeting diarization system (Evans et al., 2009). The LIA's VAD subsystem uses 12 linear frequency cepstral coefficients (LFCC) plus energy augmented by their first and second derivatives as features. The classifier is based on iterative Viterbi decoding and model adaptation applied to a two-state HMM, where one state is for speech and the other is for non-speech. Each state is initialized with a 32-component GMM trained on separate data using expectation maximization (EM)/maximum likelihood (ML) algorithm and state transition probabilities are fixed to 0.5. Also, some state duration rules are applied in order to refine the speech/non-speech segmentation.

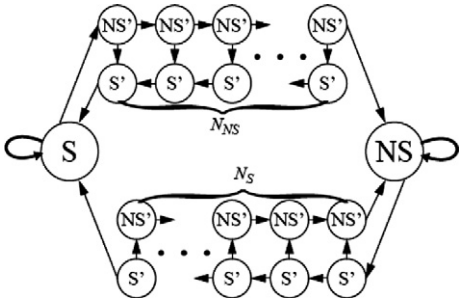

Fig. 2. Structure of ICSI-SRI HMM-based speech/non-speech detector (Pfau and Ellis, 2001).

6. Speaker segmentation

6.1. Segmentation evaluation

First of all, we introduce the segmentation evaluation measures. Two most common metrics for speaker segmentation evaluation, namely recall (RCL) and precision (PRC), are defined as follows:

- Recall: percentage of truly detected speaker boundaries (RCL).
- Precision: percentage of candidate speaker boundaries which are the actual speaker change points (PRC).

$$RCL = \frac{\text{Number of truly detected speaker boundaries}}{\text{Number of actual speaker boundaries}} \quad (1)$$

$$PRC = \frac{\text{Number of truly detected speaker boundaries}}{\text{Number of detected speaker boundaries}} \quad (2)$$

To consider the trade-off between these two metrics, the harmonic mean of these two metrics is used as the total evaluation criterion:

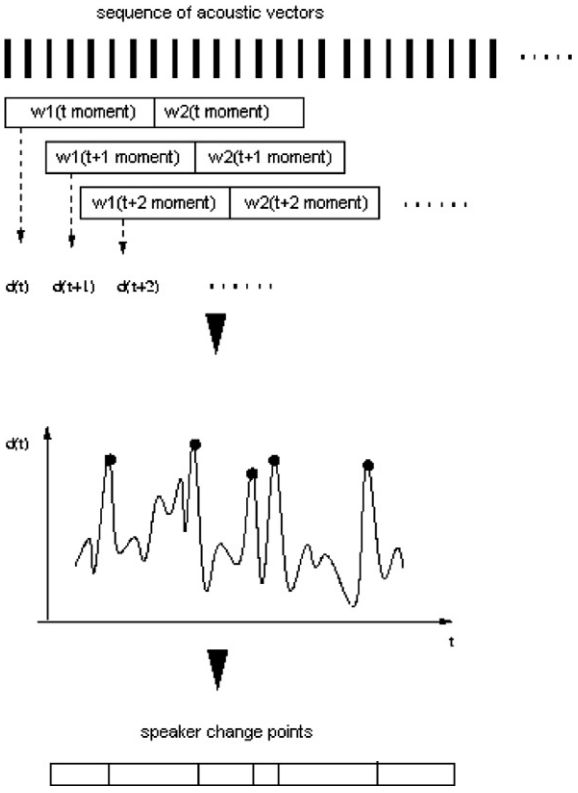

Fig. 3. Distance based speaker segmentation and change point detection (Moraru and Besacier, 2003).

$$F = \frac{2^{*}PRC^{*}RCL}{PRC + RCL} \tag{3}$$

The higher RCL, PRC and F are, the better is the performance. On the other hand, one may use the false alarm rate (FAR) and the miss detection rate (MDR) defined as:

$$FAR = \frac{\text{Number of false alarms}}{\text{Number of detected speaker boundaries}} \tag{4}$$

$$MDR = \frac{\text{Number of miss detections}}{\text{Number of actual speaker boundaries}} \tag{5}$$

A false alarm occurs when a speaker change point is detected, but it does not exist. A miss detection occurs when an existing speaker change point is not detected by the algorithm. In contrast to the first measures, the best performance is achieved when the FAR and MDR measures are the lowest.

6.2. Speaker segmentation approaches

Speaker segmentation methods are very diverse and include a huge amount of previous works on speaker based indexing. Segmentation techniques are categorized in one of the following groups:

1. ***Silence detection based methods***: Some of the speaker segmentation techniques are based on silence detection in speech signal. In these methods, it is supposed that there exists a silence region between the utterances of each two speakers. In principle, these approaches depend on thresholding the short-term energy. However, the accuracy of these techniques is poor (Kemp et al., 2000).Systems falling into this category are energy-based and decoder-based systems. The energy-based systems use an energy detector to find the points where a speaker change point most probably exists. A threshold is usually used to determine the potential silence (Kemp et al., 2000; Wactlar et al., 1996; Nishida and Kawahara, 2003).In contrast, decoder-guided segmenters run a full recognition system and obtain the change points from the detected silence locations (Liu and Kubala, 1999; Kubala et al., 1997; Woodland et al., 1997; Lopez and Ellis, 2000; Wegman et al., 1997). They normally constrain the minimum duration of the silence segments to reduce false alarms. Some of these systems use extra information from the decoder, such as gender (Tranter et al., 2004) or bandwidth (Hain et al., 1998).However, there is not a clear relationship between the existence of a silence in a recording and a change of speaker. Therefore, such systems usually take the detected change points as hypothetic speaker change points, and then verify the candidate speaker boundaries using other techniques.
2. ***Model-based segmentation***: In this class of methods, a set of models is derived and trained for different speaker classes from a training corpus. The audio stream is then classified by ML selection using these models (Gauvain et al., 1998; Kemp et al., 2000; Kubala et al., 1997; Bakis et al., 1997; Sankar et al., 1998). The boundaries between models will be the segmentation change points. Therefore, prior knowledge is a prerequisite to initialize the speaker models. Starting from the less complicated case, a universal background model (UBM) is used in (Barras et al., 2006; Wu et al., 2003; Wu et al., 2003) as the generic speaker model. Alternatively, a set of predetermined speaker models that are built by sampling the speaker space is used in (Kwon and Narayanan, 2004) which are called sample speaker models (SSM). A more complicated technique is to use anchor models, where a speaker utterance is projected onto a space of reference speakers (Collet et al., 2003). Pre-trained speaker models can be created by means of HMMs (Meignier et al., 2006; Kim et al., 2005; Pellom and Hansen, 1998; Ajmera and Wooters, 2003) or support vector machines (SVMs) (Mesgarani et al., 2004; Arias et al., 2005). Model-based segmentation algorithms tend to achieve a moderate recall rate at a high precision rate.
3. ***Distance based methods***: This kind of methods are the most common and widely used segmentation techniques and are appropriate for online indexing tasks. In these methods a distance metric between every two consecutive analysis segment, is used as a decision measure for determining the change point (Sinha et al., 2005; Siegler et al., 1997; Gauvain et al., 1998). An illustration of this group of approaches is shown in Fig. 3. These methods require a detection threshold to be empirically tuned for changes in audio type and features. Distance based methods do not guarantee the performance in different environments and applications, and tuning their parameters is one of the main obstacles for achieving a robust segmentation performance. Metric based methods do not require any prior knowledge on the number of speakers, their identities, or the signal characteristics. Metric based segmentation algorithms generally yield a high recall rate at a moderate precision rate.
4. ***Phonelword level segmentation***: Alternatively, or in addition to the above methods, a word or phone decoding step may be used to help finding speaker change points (Tranter et al., 2004). These approaches over-segment the speech data and require some additional processes to form usable segments.
5. ***Hybrid speaker segmentation***: Hybrid algorithms combine distance and model based techniques. Usually, distance-based segmentation is used initially to pre-segment the input audio signal. The resulting segments are then used to create a set of speaker models. Then, model-based re-segmentation yields a more refined segmentation.

6.2.1. Distance based segmentation

6.2.1.1. BIC segmentation. The most popular criterion for speaker segmentation is Bayesian information criterion

(BIC) (Tritschler and Gopinath, 1999; Sian Cheng and min Wang, 2003; Ajmera et al., 2004; Kotti et al., 2006; Kotti et al., 2006; Delacourt and Wellekens, 2000; Chen and Gopalakrishnan, 1998; Zhou and Hansen, 2005; Cettolo and Vescovi, 2003; Vescovi et al., 2003; Cettolo et al., 2005). BIC speaker segmentation was originally introduced in Chen and Gopalakrishnan (1998) and is derived from the generalized likelihood ratio (GLR) (Sian Cheng and min Wang, 2003).

BIC is an optimal Bayesian model selection criterion used to decide which of the models represents data samples best. This technique searches for change points within a window using a penalized likelihood ratio test of whether the data in the window is better modeled by a single distribution (no change point) or two different distributions (change point). Suppose, the samples x_i are d-dimensional feature vectors. Assuming that two neighboring analysis windows X and Y are located around time t_j, the problem is to decide whether or not a speaker change point occurs at t_j. Let $Z = X\ U\ Y$. The problem is formulated as a statistical test between two hypotheses. Under H_0, there is no speaker change point at time t_j. The data samples in Z are modeled by a multivariate Gaussian pdf, θ_Z. The log likelihood L_0 is calculated as:

$$L_0 = \sum_{i=1}^{n_X} \log p(x_i|\theta_Z) + \sum_{i=1}^{n_Y} \log p(y_i|\theta_Z) \tag{6}$$

where n_X and n_Y are the numbers of data samples in analysis windows X and Y, respectively. Under H_1, a speaker change point exists at time t_j. The analysis windows X and Y are modeled by two multivariate Gaussian densities, which are denoted by θ_X and θ_Y, respectively. Then, the log likelihood L_1 is obtained by:

$$L_1 = \sum_{i=1}^{n_X} \log p(x_i|\theta_X) + \sum_{i=1}^{n_Y} \log p(y_i|\theta_Y) \tag{7}$$

The dissimilarity between the two neighboring analysis windows X and Y is estimated by Δ BIC criterion defined as:

$$\Delta BIC = L_1 - L_0 - \frac{\lambda}{2}\left(d + \frac{d(d+1)}{2}\right)\log n_Z \tag{8}$$

where $n_Z = n_X + n_Y$ is the number of frames in analysis window Z and λ is a penalty factor. If $\Delta BIC > 0$, a local maximum of ΔBIC is found and time t_j is considered to be a speaker change point. If $\Delta BIC < 0$, there is no speaker change point at time t_j.

In BIC-based speaker segmentation, the choice of the analysis window size n_Z is of great importance. If n_Z is too large, it may contain more than one speaker changes and consequently yield a high number of miss detections. On the other hand, if $n_Z$is too short, the lack of data will cause poor model estimation (Zhou and Hansen, 2005), and poor segmentation accuracy.

Several implementations using BIC as segmentation metric have been proposed. Initially, Chen and Gopalakrishnan (1998) proposed a multiple changing point detection algorithm in two passes, and later (Tritschler and Gopinath, 1999; Sian Cheng and min Wang, 2003; Cettolo and Vescovi, 2003; Vescovi et al., 2003; Sivakumaran et al., 2001; Lu and Zhang, 2002) followed with one or two-pass algorithms. They all propose a system using a growing window with inner variable length analysis segments to find the change points.

In the formulation of BIC by Schwarz (1978), the numbers of acoustic vectors available to train the model were supposed to be infinite. In real applications, this becomes a problem when there is a big mismatch between the length of the two adjacent windows or clusters being compared. The penalty factor, λ, was introduced to adjust the penalty effect on the comparison. In speaker segmentation, λ serves as a threshold. The lower λ is, the larger PRC and the lower RCL are.

Penalty factor, λ, needs to be tuned to the data and therefore its correct setting has been the subject of previous studies. Some approaches propose ways to automatically select λ (Tritschler and Gopinath, 1999; Delacourt and Wellekens, 2000; Lopez and Ellis, 2000; Delacourt et al., 1999; Mori and Nakagawa, 2001; Vandecatseye et al., 2004). In Ajmera et al. (2004) a GMM is used for each of the models (θ_Z, θ_X and θ_Y) and by building the model θ_Z with the sum of models θ_X and θ_Y complexities, it cancels out the penalty term. The result is equivalent to the GLR metric. The number of parameters used to model the data in both hypotheses is forced to be the same, so that the likelihoods are directly comparable. As a result, the criterion is expected to be robust to changes in data conditions.

BIC change detection is costly. In Tritschler and Gopinath (1999), in order to improve the algorithm efficiency and allow for real-time implementation, a small analysis window, n_Z, is considered and if no speaker change point is found in the current window, the window size is increased by Δn_Z frames. These steps are repeated until a speaker change point is found or until the analysis window has reached a maximum size. Also to speed up computations, BIC is applied only to selected time instants. For example, BIC test is not performed at the borders of analysis windows and BIC computation is ignored at the beginning of large analysis windows. In (Cettolo and Vescovi, 2003; Vescovi et al., 2003; Sivakumaran et al., 2001) speed-ups are performed in computing the mean and variances of the models. In (Roch and Cheng, 2004) a MAP-adapted version of the models is presented, which allows for shorter speaker change points to be found.

Even with the efforts to speed up the processing of BIC, it is computationally more intensive than other metrics, but its good performance has kept it as the algorithm of choice in many applications. To solve the above problems, two pass algorithms are proposed in some papers. In these methods, a segmentation using another distance metric is performed and then the candidate change points are verified using BIC criterion. Pre-segmentation improves BIC segmentation accuracy (Sian Cheng and min Wang, 2003;

Cheng and Wang, 2004; Kotti et al., 2006; Delacourt and Wellekens, 2000; Zhou and Hansen, 2005; Delphine, 2010). Metrics such as Hotelling's T2 distance (Tranter et al., 2004; Zhou and Hansen, 2000), KL2 distance (Lu and Zhang, 2002), normalized GLR (also called normalized log likelihood ratio (NLLR)) (Vandecatseye et al., 2004) and weighted squared Euclidean distance (Kwon and Narayanan, 2002) have been previously proposed in this direction.

6.2.1.2. Other distance metrics. A wide variety of distance metrics could also be used for metric based speaker segmentation. Two commonly used metrics are the Kullback–Leibler (KL) divergence (Lu and Zhang, 2005; Delacourt and Wellekens, 2000; Harb and Chen, 2006; Hung et al., 2000) and the Gaussian divergence (also known as symmetric Kullback–Leibler-2 divergence) (Barras et al., 2006). The KL and KL2 distances (Siegler et al., 1997; Hung et al., 2000) are good choices due to their fast computation and acceptable results. In (Delacourt and Wellekens, 2000) the KL2 distance is considered as the first step for speaker change detection. In (Zochova and Radova, 2005) KL2 is used again in an improved version of the previous algorithm. In (Hung et al., 2000) the MFCC acoustic vectors are initially processed via principle component analysis (PCA) dimensionality reduction for each sliding segments and then Mahalanobis, KL and Bhattacharyya distances are used to determine if there is a change point.

Divergence shape distance (DSD) is another popular distance metric for speaker segmentation (Kim et al., 2005). The DSD is also used in (Lu and Zhang, 2002) as a first step of a two step segmentation system, using BIC on the refinement step. In (Lu and Zhang, 2002) some speed-ups are proposed to make the previous system real-time.

Also, cross-BIC (XBIC) is introduced in (Anguera et al., 2005; Anguera and Hernando, 2004) for speaker segmentation, which derives a distance between two adjacent segments by cross likelihood evaluation, inspired from the BIC distance. Authors of Malegaonkar et al. (2006) propose a similar metric and study different likelihood normalization techniques to make the metric more robust, achieving better results than BIC for speaker segmentation.

A newly introduced distance measure known as information change rate (ICR) (Han and Narayanan, 2008) or entropy can be also used to characterize the similarity of two neighboring speech segments. The ICR determines the change in information that would be obtained by merging any two speech segments under consideration and can thus be used for speaker segmentation. Unlike the measures above, the ICR similarity is not based on a model of each segment but, instead, on the distance between segments in a space of relevance variables, with maximum mutual information or minimum entropy. One suitable space comes from GMM component parameters (Vijayasenan et al., 2007). The ICR approach is computationally efficient and, in (Han and Narayanan, 2008), ICR is shown to be more robust to data source variation than a BIC-based distance.

In addition, one could perform segmentation with various metrics and/or various classifiers, and then fuse the individual results. In general, fusion offers many advantages, such as increasing the reliability, robustness, and survivability of a system (Graff, 2001). In (Zdansky, 2006), a speaker segmentation method based on the metrics such as ML, Informational and BIC is proposed. In (Omar et al., 2005) the CuSum distance (Basseville and Nikiforov, 1993), the Kolmogorov–Smirnov test (Deshayes and Picard, 1986) and BIC are first used independently to find candidate change points and then fused at likelihood level to assert those changes. In (Hung et al., 2000), the Malalanobis and Bhattacharyya distances (Campbell, 1997) are used in comparison to the KL distance for speaker change detection.

6.2.1.3. Threshold determination. All of metric-based techniques compute a function whose maxima/minima should be compared with a threshold in order to determine change points. In many cases such threshold is defined empirically given a development set. This procedure leads to a threshold which is dependent on the development set data. In the area of speaker segmentation and clustering, some publications propose automatic ways to define appropriate thresholds. In (Lu and Zhang, 2002; Lu and Zhang, 2002; Lu et al., 2002) an adaptive threshold is made dependent on the Q previous distances as follows:

$$Th_i = \alpha \frac{1}{Q} \sum_{q=0}^{Q} Dist(i-q-1, i-q) \tag{9}$$

where α is an amplification coefficient (usually set close to 1) and $Dist(i,j)$ is the distance between the ith and the jth sliding segment. The same adaptive threshold is used in (Wu et al., 2003; Wu et al., 2003; Wu et al., 2003) to evaluate

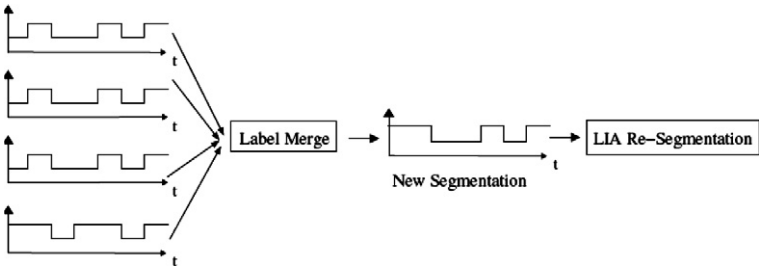

Fig. 4. Merging strategy of CLIPS and LIA system.

the difference between the local maxima and the neighboring minima distance points. In (Rougui et al., 2006) a dynamic threshold is defined for comparing speaker clusters where a population of clusters is used to decide on the threshold value. This threshold is defined as:

$$Th = \max(hist(Dist(\theta_i, \theta_j), \forall i \neq j) \quad (10)$$

where *hist* denotes the histogram and *Dist*() is the KL distance between two models.

6.2.2. Hybrid speaker segmentation

A hybrid speaker segmentation algorithm is developed in (Cheng and Wang, 2004). In this approach, after an initial BIC segmentation, the changes that are not found by BIC are detected in a top-down manner, i.e. through a divide-and-conquer technique. Another interesting hybrid system is introduced in (Meignier et al., 2006) where two systems are combined, namely the LIA system, which is based on HMMs, and the CLIPS system, which performs BIC-based speaker segmentation. The systems are combined with two different strategies. The first strategy, called hybridization, feeds the results of CLIPs system into the LIA system, whereas the second strategy, named merging (Fig. 4), merges preliminary results from LIA and CLIPs system and re-segmentation is performed using the LIA system.

6.3. Recall versus precision

A main concern in speaker segmentation is the relation between false alarm rate and change point miss detection. The research community tends to treat false alarms as less cumbersome than missed detections. However, such a consideration highly depends on the application. Over-segmentation caused by a high number of false alarms is easier to remedy than under-segmentation, caused by high number of miss detections (Barras et al., 2006; Lu and Zhang, 2002; Lu and Zhang, 2005; Sian Cheng and min Wang, 2003; Cheng and Wang, 2004; Delacourt and Wellekens, 2000; Zhou and Hansen, 2005; Hansen, 2005). Over-segmentation, for example, could be improved by clustering and/or merging. This explains why penalty factor in BIC segmentation is usually selected to yield a lower number of miss detections at the expense of a higher number of false alarms. Authors of Otero et al. (2010) present two strategies for reducing the false alarm rate with a minimal impact on the true speaker change detection rate. One of these strategies rejects those changes that are likely to be false alarms because of their low ΔBIC value, given a discarding probability. The other one assumes that the occurrence of changes constitutes a Poisson process, so changes will be discarded with a probability that follows a Poisson cumulative density function.

It is clear that speaker segmentation results affect speaker clustering. Especially, when a change point is missed and the data from two speakers is concatenated in a single speech segment, the speaker clustering performance highly degrades. It is crucial that the speech segments be homogeneous. This motivates over-segmentation and false alarms are considered less erroneous than miss detections.

On the other hand, since the change point detection often only provides an initial base segmentation for diarization systems, which will be clustered and often re-segmented later, being able to run the change point detection very fast is often more important than any performance degradation. In fact, some researchers found no significant performance degradation when using a simple initial uniform segmentation within their systems (Wooters et al., 2004; Meignier et al., 2005).

7. Speaker clustering

The goal of speaker clustering is to associate segments from an identical speaker together. Speaker clustering ideally produces one cluster for each speaker with all segments from a given speaker in a single cluster. Speaker indexing can be divided into online and offline categories based on processing requirements. Most of the previous indexing approaches perform the task in an offline manner with the assumption that the entire speech file is available in the time of decision making. In these techniques, the speech segments are merged in consecutive iterations of a hierarchical clustering algorithm and the final clusters corresponding to each speaker are constructed. Offline speaker indexing can be used for record keeping, but it is not suitable for real-time meeting or teleconferencing that demand online processing. This section exclusively concerns the offline approaches and the online methods are introduced and discussed in the next section.

7.1. Clustering evaluation

There are several performance metrics for speaker clustering performance evaluation.

Average cluster purity (acp) and average speaker purity (asp): Average cluster purity (acp) reduces when a cluster includes segments from two or more speakers. On the contrary, average speaker purity (asp) reduces when speech of a single speaker is split to more than one cluster. The best clustering scheme is the one which takes both factors into account. The *acp* criterion is based on cluster purity which is computed as follows:

$$p_i = \sum_{j=1}^{N_s} n_{ij}^2 / n_i^2 \quad (11)$$

where N_s is the number of speakers and n_{ij} is the total number of frames in cluster i spoken by speaker j. Also, n_i is the total number of frames in cluster i. Then the *acp* can be defined as:

$$acp = \frac{1}{N} \sum_{i=1}^{N_c} p_i n_i \quad (12)$$

where N_c is the number of final clusters. Similarly the *asp* criterion can be computed using speaker purity:

$$p_j = \sum_{i=1}^{N_c} n_{ij}^2 / n_j^2 \tag{13}$$

$$acp = \frac{1}{N} \sum_{j=1}^{N_s} p_j n_j \tag{14}$$

where n_j is the total number of frames spoken by speaker j. Higher asp means that the speaker data has not been distributed between many clusters. When the value of *acp* is higher, it means that most of the speech data in each cluster belongs to one speaker. In order to compare the accuracy of a clustering method, the geometrical mean of asp and *acp* factors is proposed as follows:

$$Acc = \sqrt{acp^* asp} \tag{15}$$

The *Acc* measure is also denoted by K in literature. The values of *asp*, *acp* and *Acc* are between zero and one. In optimal case when the clustering is done perfectly, *acp*, *asp* and *Acc* factors will be equal to 1.

Rand index: Rand index gives the probability that two randomly selected segments belong to the same speaker but are hypothesized in different clusters, or two segments are in the same cluster but come from different speakers. Rand index is defined as Liu and Kubala (2004), Jain and Dubes (1988):

$$\gamma = \frac{1}{C(N_s,2)} \left[\frac{1}{2} \sum_{i=1}^{N_c} n_i^2 + \frac{1}{2} \sum_{j=1}^{N_s} n_j^2 - \sum_{i=1}^{N_c} \sum_{j=1}^{N_s} n_{ij}^2 \right] \tag{16}$$

where $C(N_s,2)$ denotes the number of combinations of N_s segments by 2. Its value changes between 0 and 1. A perfect clustering should yield a zero Rand index (Jain and Dubes, 1988).

Misclassification rate: Given a one-to-one speaker to cluster mapping, if any segment from speaker j is not mapped to a cluster, an error is committed. Let $error_j$ denote the total number of segments uttered by speaker j that are not mapped to the corresponding cluster. The misclassification rate (MR) is defined in (Liu and Kubala, 2004) as follows:

$$MR = \frac{1}{N_s} \sum_{j=1}^{N_s} error_j \tag{17}$$

which ranges between 0 and 1. Small values of MR indicate a small probability of unmapped segments to any cluster.

7.2. Offline speaker clustering

Offline clustering algorithms require all data to be available before clustering. Therefore, offline speaker clustering should work better than online clustering, because it exploits more information. Most of offline clustering algorithms use hierarchical schemes, in which speech segments are iteratively split or merged until the optimum number of speakers is reached. Bottom-up (agglomerative) clustering approaches are those that start with the segments from the segmentation phase and converge to the final clusters via merging techniques. On the other hand, top-down algorithms start with one or very few clusters and obtain the optimum clusters via splitting procedures. Offline clustering systems mainly differ in the selection of the distance function, how clusters are merged and the stopping criterion. In both approaches, two items need to be defined:

1. A distance between clusters/segments to determine acoustic similarity. Distance metric defines the distance of a cluster from any other possible cluster. Distance metrics used for speaker clustering can be similar to those used in speaker segmentation.
2. A stopping criterion to stop the iterative merging/splitting at the optimum number of clusters. This measure should be considered because the final number of clusters is not known at the beginning of the clustering.

7.2.1. Bottom-up clustering

The most common approach in speaker clustering is agglomerative clustering which consists of the following steps:

1. Initializing clusters with speech segments.
2. Computing pair-wise distances between clusters.
3. Merging closest clusters.
4. Updating distances of remaining clusters.
5. Iterating steps 2 to 4 until stopping criterion is met.

In speaker clustering, the clusters are generally represented by a single full covariance Gaussian (Sinha et al., 2005; Reynolds and Torres-Carrasquillo, 2004; Moh et al., 2003; Nguyen et al., 2002; Barras et al., 2004), but GMMs have also been used (Wooters et al., 2004; Meignier et al., 2005), sometimes being built using MAP adaptation of a UBM for increased robustness.

7.2.1.1. Distance measure. BIC is an appropriate measure for decision on the similarity of the clusters in hierarchical speaker clustering (Cettolo and Vescovi, 2003; Delphine, 2010; Chen et al., 2002). The BIC similarity criterion compares the statistics of the two clusters being merged, x and y, with that of the resulted merged cluster, z. supposing that the clusters are modeled with Gaussian distributions, ΔBIC for merging two clusters is calculated as follows:

$$\Delta BIC = \frac{1}{2} [n_z \log(|\Sigma_z|) - n_x \log(|\Sigma_x|) - n_y \log(|\Sigma_y|)] - \lambda \left(\frac{d(d+3)}{4} \right) \log n_z \tag{18}$$

where Σ is the covariance matrix, and d is the dimension of the feature vector. If the pair of clusters is best described by a single full covariance Gaussian, the ΔBIC will be low, whereas if there are two separate distributions, implying two speakers, the ΔBIC will be high. In each clustering step,

the pair of clusters with the lowest ΔBIC is merged and the statistics are recalculated. The process is stopped when the lowest ΔBIC is greater than a specified threshold. In some previous researches modifications is proposed to the penalty term (Tranter et al., 2004; Tritschler and Gopinath, 1999; Cettolo and Vescovi, 2003; Chen et al., 2002; Meinedo and Neto, 2003). The system described in (Wooters et al., 2004; Ajmera and Wooters, 2003) removes the need for tuning the penalty factor on development data, by ensuring that the number of parameters in the merged and separate distributions are equal. Alternatives to the penalty term, such as using a constant (Tranter et al., 2004), the weighted sum of the number of clusters and the number of segments (Gauvain et al., 1998), or a penalized determinant of the within cluster dispersion matrix (Liu and Kubala, 2003; Jin et al., 1997) have also been proposed.

Introduction of Local BIC and its significant improvement in the speaker diarization accuracy (Stafylakis et al., 2009), showed that the original formulation of BIC in (Chen and Gopalakrishnan, 1998), (i.e. the Global BIC) is far from being optimal for the speaker diarization task. Local BIC is an autonomous pair-wise dissimilarity measure, i.e. the corresponding ΔBIC formula is completely defined by the sufficient statistics of the two clusters being examined and their sizes. To combine the strengths of the two approaches, (Stafylakis et al., 2009) proposes a new variant, the segmental BIC. The idea is to redefine the priors of the BIC, so that the corresponding ΔBIC becomes autonomous. The results showed that the segmental BIC is at least comparable to the local BIC and superior to the global BIC, especially in applications where the purity of the clusters is more important than their coverage. Stafylakis et al. (2010) demonstrates that the results can severely be improved by retaining the main idea of the Segmental BIC and posing a stricter penalty term.

Also, other distance metrics can be employed in agglomerative speaker clustering. In (Sankar et al., 1995; Heck and Sankar, 1997) the symmetric relative entropy distance (Juang and Rabiner, 1985) is used for speaker clustering. From the other metrics which are used in cluster similarity computation in speaker clustering, the Gish distance (Jin et al., 1997), the KL distance (Siegler et al., 1997), KL2 distance (Zhou and Hansen, 2000), and cross likelihood ratio (CLR) (Fernandez et al., 2009) can be mentioned.

7.2.1.2. Model-based speaker clustering. In some state-of-the-art approaches, the speech clusters are modeled at each step and then the relative distance between these models is used as the clustering measure. Many approaches based on GMMs have been proposed in this category. Authors of Solomonoff et al. (1998) propose a method for clustering speakers based on GMMs. Each speaker is modeled by a GMM. The algorithm has three stages. The first stage aims at computing some distance measures between each pair of speech segments, such as the cross entropy which is defined as follows (Solomonoff et al., 1998):

$$Dist_{CE}(s_0, s_1) = \log\frac{p(s_0|\theta_{s_0})}{p(s_0|\theta_{s_1})} + \log\frac{p(s_1|\theta_{s_1})}{p(s_1|\theta_{s_0})} \quad (19)$$

where θ_s denotes the model trained on segment s. A tree of clusters is created at the second stage, where each segment forms its own cluster at the beginning, and clusters are merged recursively. The merging scheme is continued until the estimated cluster purity is maximized.

The algorithm proposed in Jin et al. (1997) assumes that each segment can be modeled by a Gaussian distribution. A distance matrix is built based on Gaussian models of acoustic segments. The hierarchical clustering procedure takes the distance matrix as input and merges clusters together until one sufficiently large cluster is formed. The output is a tree of clusters, which can be pruned for any given number N_c. Finally, model selection is conducted by employing a criterion that penalizes too many clusters.

Speaker modeling and comparison can also be performed in reference space instead of original feature space. Tsai et al. propose a speaker clustering method which is based on the voice characteristic reference space (Tsai et al., 2004). The reference space aims at representing some generic characteristics of speaker voices derived through training. The speech features are projected onto a reference space to be clustered. The propositions for the construction of the reference space (Tsai et al., 2004) include the utterance Gaussian mixture modeling, the utterance vector clustering, and universal Gaussian mixture modeling followed by model adaptation. Having constructed the reference space, each segment s_i is mapped onto a K dimensional projection vector V_i, where v_i denotes how much segment s_i can be characterized by the basis vector i. The similarity between any two segments s_i and s_j is computed using the cosine similarity measure:

$$CSM(s_i, s_j) = V_i^T V_j / \|V\|_i \|V_j\| \quad (20)$$

The segments which are similar enough are grouped into a cluster. In (Biatov and Köhler, 2006), speaker characteristic vectors are used in the clustering scheme.

A further method described in (Reynolds and Torres-Carrasquillo, 2004; Akita and Kawahara, 2003) uses anchor speakers. A set of anchor models is applied to map segments into a vector space (Fig. 5), then a Euclidean distance metric and a stopping criterion are used, but the overall clustering framework remains the same. The anchor models can be built by adapting a UBM model to the test data segments themselves, thus making the system portable to different domains.

In literature, speaker subspace learning has gain popularity. The approach that represents an arbitrary utterance or speaker as a linear combination of a set of basis voices based on PCA is eigen-voice approach (Kuhn et al., 1998; Kuhn et al., 2000). The eigen-voice approach has been demonstrated to be successful for both speaker recognition (Thyes et al., 2000) and diarization (Castaldo et al., 2008), and to outperform algorithms directly working in the original speaker space. Short speech segments are the main

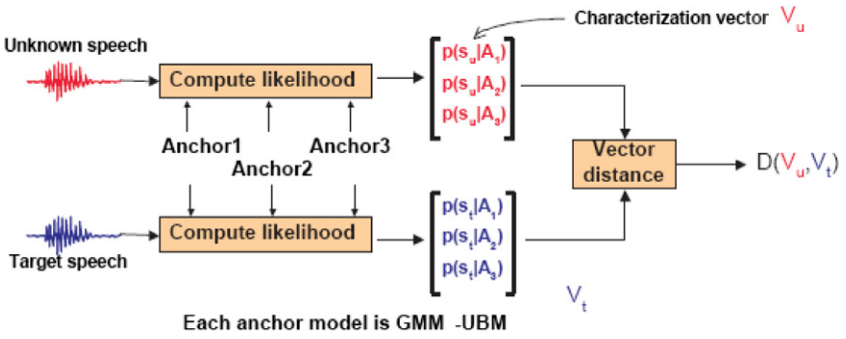

Fig. 5. Anchor model system (Sturim et al., 2001).

source of error in speaker clustering algorithms. These segments may not be modeled efficiently especially when the benchmark system is based on GMM speaker models. In (Moattar and Homayounpour, 2009) eigen-voice speaker adaptation methods are studied for enhancing speaker modeling and clustering performance, especially in the presence of short speech segments.

In Chu et al. (2009) the fisher-voice approach is proposed for speaker clustering. The fisher-voice approach is based on linear discriminant analysis (LDA) and provides an optimized low-dimensional representation of utterances or speakers with focus on the most discriminative basis voices. In this approach, a training set is used to train a UBM. The UBM is then MAP adapted to give a GMM for each training utterance. Any utterance in the training set can be represented by a GMM mean supervector. Next, LDA (or PCA) is performed on the training set in the GMM mean supervector space to derive the fisher-voice (or eigen-voice) space. Given a test utterance, the UBM is MAP adapted to form a GMM mean supervector for this test utterance, which is further projected onto the fisher-voice (or eigen-voice) space. Finally, hierarchical agglomerative clustering technique based on the Euclidean distance metric and the Ward's linkage method (Ward, 1963) is used to perform unsupervised clustering in the fisher-voice (or eigen-voice) space. The experiments show that the fisher-voice approach significantly outperforms the eigen-voice approach.

7.2.2. Top-down clustering review

Top-down clustering is computationally much more expensive since all the possible splitting points should be compared. Hence, There are fewer divisive clustering approaches than the agglomerative methods. In (Johnson and Woodland, 1998) a top-down clustering method is proposed for speaker clustering towards ASR, and in (Tranter et al., 2004; Johnson, 1999) it is applied to speaker diarization. The algorithm splits the data iteratively into four sub-clusters and allows for merging clusters that are very similar to each other. Authors of Johnson and Woodland (1998) propose two different implementations of the algorithm. On one hand, it proposes a maximum likelihood linear regression (MLLR) likelihood optimization technique to obtain resulting clusters. On the other hand, it proposes the arithmetic harmonic sphericity (AHS) metric (Bimbot and Mathan, 1993) to assign speech segments to the created sub-clusters at each stage, and uses a minimum occupancy stopping criterion. The AHS is defined for single Gaussian models as:

$$AHS(s_1,s_2)=\log\left[tr(\Sigma_{s_2}\Sigma_{s_1}^{-1}).tr(\Sigma_{s_1}\Sigma_{s_2}^{-1})\right]-2\log(d) \qquad (21)$$

where $tr()$ is the trace function.

7.2.3. Combinational clustering

Given that both Top-down and bottom-up techniques could complement each other, some people have proposed systems that can combine multiple systems and obtain an improved speaker diarization. Hain et al. (1998) compared both clustering techniques in broadcast news transcription. On one hand, bottom up clustering uses a divergence-like distance measure and a minimum cluster feature count as stopping criterion. On the other hand, top-down clustering uses the arithmetic harmonic sphericity distance and also the cluster count as stopping criterion.

In (Tranter, 2005) a cluster voting algorithm, as shown in Fig. 6, is presented to allow diarization output improvement by merging two different speaker diarization systems. Tests are performed using two top-down and two bottom-up systems. In (Moraru, 2004; Fredouille et al., 2004) two different combination approaches are presented to combine top-down and bottom-up outputs for broadcast news and meetings speech diarization. A technique called hybridization proposes one system as initialization to the second system. The second technique called Fusion proposes a matching of common resulting segments followed by a re-segmentation of the data.

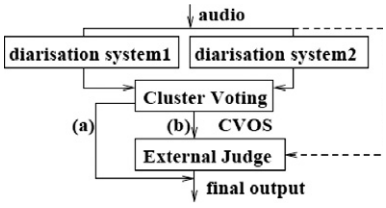

Fig. 6. Cluster voting architecture (Tranter, 2005).

7.2.4. Evolutionary HMM-based speaker clustering

Ajmera et al. (2002), Ajmera et al. (2002) propose an HMM-based speaker clustering algorithm. Each state of the HMM represents a cluster and each cluster is modeled by a GMM. The initialization of the pdfs is done using the *K*-means algorithm. The technique starts with over-clustering the data in order to reduce the probability that different speakers are clustered into one class. Afterwards, the segmentation is performed using the Viterbi algorithm in each cluster. The next step is to reduce the number of clusters by merging. The clusters are merged according to a likelihood ratio distance measure. The new class is represented by another GMM. The segmentation is re-estimated with the new HMM having one cluster less and the likelihood of the data based on this segmentation is calculated. The likelihood increases, if the data in the two clusters to be merged are from the same speaker. In contrary, it decreases if the clusters to be merged have data from different speakers. The merging process stops when the likelihood does not decrease any more.

The algorithm proposed in (Ajmera and Wooters, 2003) uses HMMs, agglomerative clustering, and BIC. The number of states in the HMM is equal to the initial number of clusters. Each state is composed of a set of sub-states. The sub-states impose a minimum duration on the model. Each state of the HMM is a cluster and is expected to represent a single speaker. The HMM model of the clusters is illustrated in Fig. 7. The algorithm starts with over-clustering the data. The first step is to initialize the parameters of HMM. The initialization is performed using a uniform segmentation and estimating the parameters of the cluster GMMs over these segments. In the E-step, a segmentation of the data is obtained to maximize the likelihood of the data, given the parameters of the GMM. In the M-step, the Parameters of the GMM are re-estimated based on the new segmentation. The final step is cluster merging. BIC can be used as a merging criterion.

7.2.5. Stopping criterion

A difficult problem in speaker clustering is the estimation of the number of clusters to be created. Ideally, the number of clusters, N_c, should be equal to the number of speakers, N_s. However, N_c, is generally greater than or equal to N_s. Authors of Voitovetsky et al. (1998) propose a validity criterion for automatically estimating the number of speakers in self organizing map (SOM)-based speaker clustering (Voitovetsky et al., 1997; Lapidot and Guterman, 2001; Lapidot, 2002). The data are divided into short segments. Each segment is considered to be long enough to enable determining speakers' identity. Initially, speech segments are randomly and equally divided between N_s models. The speech segments are clustered into N_s speakers by performing competition between the SOMs. Multiple iterations are allowed during training. After each iteration, the data is regrouped. The training process is applied again to the new partition until the partitions remain unchanged or their difference between two consecutive iterations is less than a threshold value. For efficient clustering, the intra-cluster distance should be relatively small, while the inter-cluster distance should be relatively large. Thus, the proposed criterion defines the validity of a given partition to be proportional to the ratio of the intra-cluster distance over the inter-cluster distance. Ben-Harush et al. (2008) employs weighted segmental k-means clustering algorithm prior to competitive based learning. This approach focuses on the initial assignment of feature vectors to N_s clusters. Initial assignments form the bases for the iterative competitive learning based clustering system.

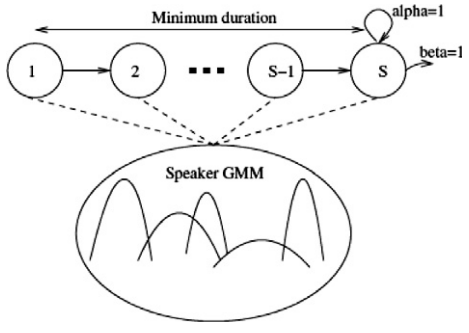

Fig. 7. HMM model with minimum duration for speaker clustering (Ajmera and Wooters, 2003).

Another algorithm for the automatic estimation of the number of clusters is based on the BIC (Chen and Gopalakrishnan, 1998). Let $C_{current}$ be the current clustering with N_c clusters and n be the number of the data samples to be clustered. Each cluster is modeled by a Gaussian distribution. BIC for $C_{current}$ is defined as:

$$BIC(C_{current}) = \sum_{i=1}^{N_c} -\frac{1}{2} n_i \log|\Sigma_i| - N_c n\left(d + \frac{d(d+1)}{2}\right) \quad (22)$$

where Σ_i is the covariance matrix of the ith cluster and $|.|$ denotes the matrix determinant. The clustering which maximizes BIC is chosen. However, it is costly to search globally for the best BIC value. In hierarchical clustering methods, it is possible to optimize the BIC in a greedy fashion (Barras et al., 2006; Tranter and Reynolds, 2006; Chen and Gopalakrishnan, 1998).

Regardless of the employed clustering method, the stopping criterion is critical for good performance. Under-clustering, fragments speaker data over several clusters, while over-clustering produces contaminated clusters containing speech from several speakers. For speaker indexing of speech signal, both under-clustering and over-clustering are suboptimal. However, when using clustering outputs to assist in speaker adaptation of ASR systems, under-clustering is suitable. Also, over-clustering may be advantageous in aggregating speech from similar speakers or acoustic environments.

7.2.6. Clustering initialization

Most, if not all, of agglomerative hierarchical clustering approaches require a certain level of manual tuning of the

initialization parameters, including the initial amount of clusters, k, and the initial number of Gaussians per cluster, g. The robustness of these systems depends heavily on the manual tuning of the above mentioned parameters. Previous work in initialization methods for speaker diarization has concentrated on adapting the initialization parameters and performing non-uniform initialization strategies. In (Anguera, 2006), the cluster complexity ratio (CCR) is used to adapt initialization parameters k and g and in (Leeuwen and Konecny, 2008), the constant seconds per Gaussian (CSPG) is used to adapt g and to initialize the system. In (Ajmera and Wooters, 2003), a uniform initialization is compared to a K-means initialization and it is claimed that the type of initialization does not have significant impact on the result. Imseng and Friedland (2010) presents a novel adaptive initialization scheme that can be applied to most state-of-the-art speaker diarization algorithms. The initialization method is a combination of the recently proposed adaptive seconds per Gaussian (ASPG) method and a new pre-clustering and number of initial clusters estimation method based on prosodic features.

8. Speaker tracking and online diarization

In some applications, it is important to produce speaker labels immediately without collecting all of the potential data from a particular scenario, such as real-time captioning of broadcast news. This constraint prevents the standard hierarchical clustering techniques, and instead requires the clustering to be performed sequentially or online.

One of the main technical differences between offline and online speaker indexing is the feasibility of multi-pass processing over the same data. In offline indexing, it is possible to use various speaker indexing algorithms in each iteration. However, in online tracking the speech stream is gradually fed to the system, and therefore, there is not much information on the future data. Therefore, traditional clustering methods which are mainly used in offline indexing cannot be applied in this context.

In online speaker indexing, we are limited to make all decisions using only the current and previously seen speech data. Furthermore, since the models of speakers are not available *a priori* for indexing, we need to create and update them on the fly. This leads to a number of challenges. Under these circumstances of sequential learning, data is not sufficient to build a speaker model initially. A roughly built model is apt to cause decision errors due to small initial amount of data and model initialization problems. Without good initial models for speaker indexing, we cannot effectively build/update speaker models sequentially and incrementally. A block diagram of online speaker indexing is shown in Fig. 8.

The key issue here is to find a method for alleviating the model initialization problem. To address this problem, generic models are proposed as side information (Kwon and Narayanan, 2003a). Generic models are also called as reference speaker models in the literature. These models can help initialize the models and also provide a measure for computing the similarity between incoming segments. This idea is built on the hypothesis that independent speech data corpus can help initialize a model set for the unsupervised indexing. The generic model set is predetermined by training. Note that the speakers in the training data are independent from speakers in the test data. In other words, the reference model set can be used for initializing and bootstrapping any speaker indexing process. Generic models approach is found to provide better performance than other model bootstrapping methods in unsupervised speaker indexing (Kwon and Narayanan, 2003a).

Generic speaker models are originally introduced by Kwon and Narayanan (2003a), Kwon and Narayanan (2003b), Kwon and Narayanan (2004a) for unsupervised online indexing. These references propose three different choices for generic models which include the UBM, Gender Background Models (GBM) and a set of sample speaker models (SSMs). From the mentioned approaches, SSMs had achieved the most efficient indexing solution. These approaches have exploited the GLR test for speaker segmentation and a segmentation modification algorithm called localized search algorithm (LSA) for change point modification and refinement. Also the SSM models are selected from a pool of general speaker models using the Markov Chain Monte Carlo (MCMC) sampling method.

The method proposed in (Kwon and Narayanan, 2004b) uses a similar approach as Kwon and Narayanan (2003a), Kwon and Narayanan (2003b), Kwon and Narayanan (2004a) except that it applies a tree structure to quantize the SSM models to achieve more general models with better distribution of speakers. This modification have resulted lower error rate on 2 speaker telephone conversations and broadcast news.

Whether using generic models or not, GMM model training is vulnerable against insufficient training or adaptation data. In (Nishida and Kawahara, 2003), GMM and vector quantization (VQ) models are simultaneously used for improving the tracking performance when a small amount of training data is available. Conventionally, GMM and VQ-based methods are used in speaker recognition. It is known that the recognition performance of GMM is higher than VQ but GMM cannot be estimated with a small size of

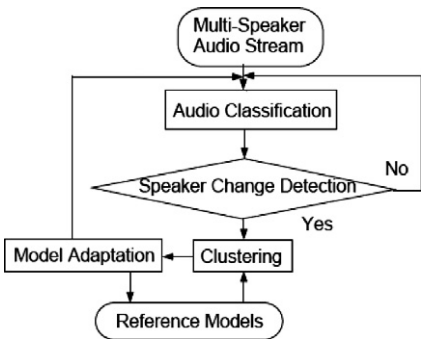

Fig. 8. Block diagram of online speaker indexing (Kwon and Narayanan, 2003b).

data. In (Nishida and Kawahara, 2003), an optimal speaker model is selected based on BIC which reflects the amount of speech data, and the speaker models are directly estimated without using an adaptation technique.

Another interesting online diarization system is introduced in (Rougui et al., 2006) which detects change points as soon as the speech becomes available and data is assigned to either speaker present in the database or a new speaker is created, according to a dynamic threshold. This assignment is performed using a KL distance between the GMM models of the current segment and the generic models. Given two models GMM_1 and GMM_2, with M_1 and M_2 Gaussian mixtures, respectively, and Gaussian weights $w_{1(i)}, i = 1, \ldots, M_1$ and $w_{2(j)}$, $j = 1, \ldots, M_2$, the KL distance between GMM_1 and GMM_2 is:

$$KL_m(GMM_1, GMM_2) = \sum_{i=1}^{M_1} w_1(i) \min_{j=1}^{M_2} KL(Norm_1(i), Norm_2(j)) \tag{23}$$

where $Norm_1(i)$ is the ith Gaussians of GMM_1.

Authors of Markov and Nakamura (2007a) describe an online speaker diarization system which has very low latency. It consists of several modules including voice activity detection, novel speaker detection, and speaker gender and identity classification. All modules share a set of GMMs representing pause, male and female speakers, and each individual speaker. During the speaker diarization process, for each speech segment it is decided whether it comes from a new speaker or from an already known speaker. In the case of a new speaker, his/her gender is identified, and then, from the corresponding gender GMM, a new GMM is spawned by copying its parameters. This GMM learned online is used to represent the new speaker. In the case of an old speaker, the corresponding GMM is again learned online. In order to prevent unlimited growth in the number of speaker models, those models that have not been selected as winners for a long period of time are deleted from the system. This allows the system to be able to perform its task indefinitely in addition to being capable of self-organization. Such functionalities are attributed to the so called Never-Ending Learning systems (Markov and Nakamura, 2007b). In (Markov and Nakamura, 2007a), a fixed global threshold is tuned on held-out dataset. However, detailed performance analysis showed that the optimal threshold depends on the number of registered speakers as well as on the speaker gender. In (Markov and Nakamura, 2008), different thresholds are used for male and female speakers and for each gender before thresholding likelihood ratio (LR), mean and variance normalization is applied. This greatly reduced the threshold dependency on the number of speakers and allowed to use fixed threshold for each gender. The LR distribution statistics are collected online and updated each time a new likelihood ratio is calculated.

In (Zamalloa et al., 2010) a low-latency speaker tracking system is presented, which deals with continuous audio streams and outputs decisions at one-second intervals, by scoring fixed-length audio segments with a set of target speaker models. A smoothing technique is explored, based on the scores of past segments, which increases the robustness of tracking decisions to local variability. The real-time speaker tracking system proposed in (Zamalloa et al., 2010) computes a detection score per target speaker and outputs a speaker detection decision for fixed-length segments. That length has been empirically set to one second, which provides relatively good time resolution and a reasonably small latency for most online speaker tracking scenarios.

Koshinaka et al. (2009) proposes an online speaker clustering method suitable for real-time applications. Using an ergodic hidden Markov model, it employs incremental learning based on a variational Bayesian framework and provides probabilistic (non-deterministic) decisions for each input utterance, directly considering the specific history of preceding utterances. This approach is applied to a generative model of speech utterances. It repeatedly estimates model parameters and a probabilistic clustering result for each input utterance, and updates those for a certain number of preceding utterances. It makes possible more robust cluster estimation and precise classification of utterances compared to conventional online methods. Online learning is accomplished on the basis of a generalized EM (GEM) algorithm (Neal and Hinton, 1998), which employs a maximum negative free energy criterion. For i.i.d. observations E and M-steps will be incrementally performed for each observation. The difference between online (GEM-based) and batch (traditional EM) algorithms is the E-step. At the E-step in the online algorithm, the probability distribution for the hidden variables is updated only for the latest observation, rather than for all observations to that point.

An evolutionary GMM training approach is proposed in (Lu and Zhang, 2005) for unsupervised speaker segmentation and tracking in real-time audio content analysis (Fig. 9). In this algorithm, LSPs, MFCCs, and pitch are extracted from the speech segments. These features are then fused in a parallel Bayesian Network. Afterwards, speaker segmentation is performed in an evolutive manner. An initial Gaussian model is estimated for each segment, and then the divergence between each two consecutive models is calculated. A potential speaker change point between two consecutive segments is detected, if there exists a local peak in the distance that exceeds a threshold. When no potential speaker change point is identified, the data of the current segment are assigned to the current speaker in order to yield a more accurate speaker model. If a potential speaker change boundary is detected, the algorithm searches the speaker models created thus far to identify the newly appeared speaker. If the speaker cannot be identified, a new speaker model is created. Otherwise, the identified speaker model is updated with the new speaker data.

In tracking and clustering algorithms, a wrong decision in segment assignment may propagate in the clustering procedure. In (Wang et al., 2007), in order to improve the clustering performance, a decision tree is built to record a history of the intermediate clustering results in the

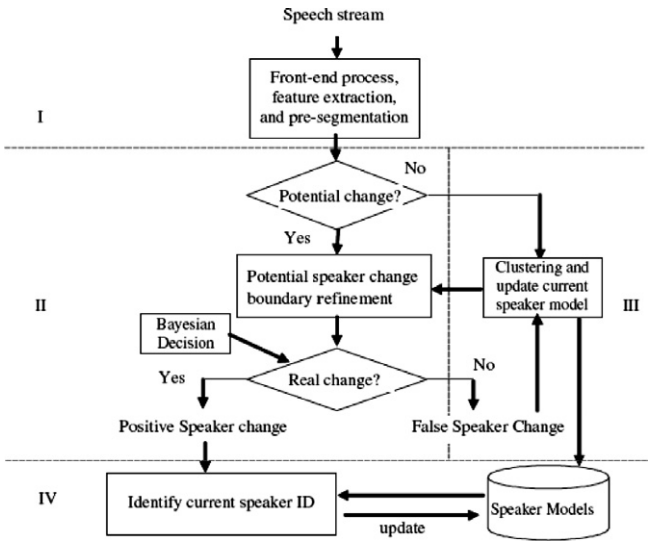

Fig. 9. The diagram of speaker change detection and speaker tracking system proposed in (Lu and Zhang, 2005), composed of four main modules: I. Frontend process. II. Speaker segmentation. III. Clustering and speaker model updating. IV. Speaker tracking.

clustering procedure. This proposition applies a tree traversing and pruning strategy to choose the best clustering path to follow. In (Wang et al., 2007), the distance metric between clusters is GLR.

9. Simultaneous segmentation and clustering

Diarization algorithms can only work well if the initial segmentation is of sufficiently high quality. Since this is rarely the case, alternative approaches combine clustering with iterative re-segmentation, hence facilitating the introduction of missing speaker turns. Most of present diarization systems thus perform both segmentation and clustering simultaneously or they perform clustering on a frame-to-cluster basis. An alternative approach to running segmentation and clustering stages separately is to use an integrated scheme. In (Moraru, 2004), LIA and CLIPS systems as two of the most recent achievements of ELISA lab in speaker indexing and tracking are introduced. LIA system uses an evolutive HMM (E-HMM) for speaker indexing. The indexing process in the system consists of segmentation and re-segmentation phases. The recording is represented by an ergodic HMM in which each state represents a speaker, and transitions model changes between speakers. The initial HMM contains only one state and represents all of the data. In each iteration, a short speech segment coming from a new speaker, is selected and used to build a new speaker model by Bayesian adaptation of a UBM. A state is then added to the HMM to reflect this new speaker, and the transition probabilities are modified accordingly. A new segmentation is then generated from a Viterbi decoding of the data with the new HMM, and each model is adapted using the new segmentation. This re-segmentation phase is repeated until no more changes are observed on the speaker labels. The process of adding new speakers is repeated until there is no gain in terms of comparable likelihood or there is no data left to form a new speaker. The main advantages of this integrated approach are to use all the information at each step and the use of speaker recognition techniques.

Bozonnet et al. (2010) modifies the E-HMM approach proposed in (Moraru, 2004) for LIA system. The main contributions of this paper are the enhancements to the top-down E-HMM speaker diarization system and presenting a new approach that delivers additional significant improvements in performance through cluster purification. Small improvements in LIA's top-down E-HMM diarization system using purification were reported in (Fredouille and Senay, 2006), but the module was later removed as subsequent developments in model initialization and speaker modeling rendered the improvements negligible.

A novel attempt at integrating a new approach to cluster purification (Nguyen et al., 2009) is applied in (Bozonnet et al., 2010). The diarization system presented in (Bozonnet et al., 2010; Nguyen et al., 2009) is initialized with 30 homogeneous clusters of uniform length and a 4-component GMM is trained on the data in each cluster. Each cluster is then split into segments of 500 ms length and the top 25% of segments which best fit the GMM are identified and marked as classified. The remaining 75% of worst-fitting segments are then gradually reassigned to their closest GMMs, K segments at a time, with iterative Viterbi decoding and adaptation, until all segments are classified. The same algorithm was presented with modifications in (Sun et al., 2009) for purification purposes and for the processing of multiple distant microphone (MDM) meeting.

In (Bozonnet et al., 2010), purification is applied after segmentation and clustering which produces a number of clusters each of which, ideally, corresponds to a single speaker. Thus, in contrast to the bottom-up approach, where the initial clustering is generally random and uniform, this cluster purification algorithm operates on clusters which should already contain a dominant speaker.

10. Multimodal diarization

Using the synchrony between audio and video, humans can locate which part of their visual input is likely the source of audio input and decide whether their perception of a speaking person is synchronized to a specific audio stream. In order to enable machines to perform synchrony detection and apply it to speaker diarization, relevant research has split the problem of synchrony detection in three steps: (1) extract features from the audio and video stream (2) measure the synchrony using these features and (3) use the synchrony measurements to detect the speaker.

Relevant research makes three implicit assumptions. First, the features extracted from the audio and video stream are assumed to contain the synchrony-related information. Second, the measure of synchrony is assumed to correspond to how likely it is that the two streams are synchronized. Third, the person appearing most synchronized to the audio stream is assumed to be the speaker.

A related class of works investigates the problem of audiovisual synchrony (Nock et al., 2003; Fisher and Darrell, 2004), where given an audio or speech signal, the aim is to localize its source in video. Another related class of works is online multimodal speaker detection (Zhang et al., 2006), where given multiple audio and video streams, one would like to know if someone is speaking and where that person is. However, such a system does not perform speaker identification and thus is unable to answer the question of "who spoke when". Furthermore, it requires a structured microphone array and an omni-directional camera system. The most related class of works would be those that investigate multi-modal speaker diarization (Vajaria et al., 2006). In (Vajaria et al., 2006), the authors used meeting data with non-frontal faces and full body motion was captured. They concatenate the audio and video features together as an early-fusion approach; incidentally, they found that the combination of audio and video does not improve their diarization performance when compared to using audio or video alone.

The authors of Zhang et al. (2006) present a multimodal speaker localization method using a specialized satellite microphone and an omni-directional camera. The work presented in (Noulas and Kros, 2007) integrates audiovisual features for on-line audiovisual speaker diarization using a dynamic Bayesian network (DBN) but tests were limited to discussions with two to three people on two short test scenarios. Another use of DBN, also called factorial HMMs (Ghahramani and Jordan, 1997), is proposed in (Noulas et al., 2009) as an audiovisual framework. Common approaches to audiovisual speaker identification involve identifying lip motion from frontal faces, e.g. (Chen and Rao, 1996; Fisher et al., 2000; Rao and Chen, 1996; Siracusa et al., 2007). Therefore, the underlying assumption is that motion from a person comes predominantly from the motion of the lower half of his/her face.

In a real scenario the subject behavior is not controlled and, consequently, the correct detection of the mouth is not always feasible. Therefore, other forms of body behavior, e.g. head gestures are used (McNeill, 2000). While there has been relatively little work on using global body movements for inferring speaking status, some studies have been carried out (Vajaria et al., 2006; Hung et al., 2008; Campbell and Suzuki, 2006) that show promising initial results.

11. Diarization evaluation

The standard performance metric of the speaker indexing and diarization systems is the RT-09 NIST rich transcription evaluation metric (Rich Transcription Meeting Recognition Evaluation Plan, 2009). To evaluate the accuracy of the indexing systems, an optimum mapping from the real speakers in the conversation to the output speakers should be found. The criterion for this mapping optimality is the percentage of the speech parts which are jointly assigned to the real speaker and the output speaker. In other words, a real speaker and an output speaker with the most common speech parts are supposed to be the same. This optimality metric is calculated for all segments and all speakers. In this mapping the following restrictions should be considered:

- Each real speaker should be mapped to at most one output speaker and each output speaker should also be mapped to at most one real speaker. If the system performance is ideal, there should exist a corresponding speaker for each real speaker.
- This mapping will be performed separately for each speech file.

Since there may exist segmentation error in reference files; 250 ms of speech signal from each sides of the change point is excluded from the evaluations. The total speaker indexing error is computed as the percentage of speech parts which are not assigned to the true speaker:

$$Error_{SpkrSeg} = \frac{\sum_{allsegs}\{dur(seg) * (\max(N_{Real}(seg), N_{Out}(seg)) - N_{Correct}(seg))\}}{\sum_{allsegs}\{dur(seg).N_{Ref}(seg))\}} \quad (24)$$

In the above equation, for each speech segment *seg*, the following values should be computed:

$Dur(seg)$: Speech segment duration,
$Num_{Real}(seg)$: The number of real speakers in the speech segment,
$Num_{Out}(seg)$: The number of output speakers in the speech segment,
$Num_{Correct}(seg)$: The number of real speakers which have participated in the speech segment and the corresponding speaker is recognized in the segment.

As mentioned above, this metric is the standard metric used in NIST rich transcription for evaluating the performance of speaker detection systems and approaches. Since in this metric, 250 ms of speech signal from each end of a speaker's utterance is removed prior to error computation, this metric is more in favor in shorter utterances. Also, since the above error rate is time-weighted, it determines little importance for the diarization quality of speakers whose overall speaking time is small.

12. Proposed framework

In this section, the most common approaches previously mentioned for speaker indexing and diarization are applied in a unified framework to provide low latency online speaker indexing for various domains. The main goal of the proposed approach is to offer a set of applicable measures and algorithms for domain independent speaker indexing which satisfies the online requirements as well. To achieve the mentioned goals, the proposed framework employs distance metric based speaker segmentation which, as mentioned in Section 6 is the best choice for online change point detection. On the other hand to improve the recall and precision of the change point

detection procedure, a voting approach of different segmentation measures and algorithms is applied. Also, to facilitate online tracking, the generic models idea as introduced in (Kwon and Narayanan, 2003a,b, 2004a,b; Markov and Nakamura, 2007a) is applied. To enrich the generic models so that these models represent the speaker space better, the model creation and adaptation approaches in eigen-voice space (Kuhn et al., 2000; Huang et al., 2004) are proposed and applied in this framework. Also, we propose an index structure on the generic models and detected speaker models to speed up the model comparison and retrieval in the tracking procedure. The other proposition which is proved to be effective in improving the robustness of diarization systems (Anguera et al., 2006) is frame purification, which is also applied in the proposed framework.

The proposed diarization framework consists of the following steps. These steps are also illustrated in Fig. 10. The proposed framework starts from two different sides. From one side which is illustrated in Fig. 10(a), the generic speaker models are trained and the reference models are selected.

Building generic models: The proposed framework is based on the GMM modeling of the speakers. To build the generic models we can use the ML criterion and train the model from the beginning or as illustrated in Fig. 10(a), adapt a UBM model trained on an arbitrary speech database, using an adaptation algorithm. Adaptation may be more suitable for building generic models. Since the resulted models represent the speaker specific characteristics, they inherit the characteristics of the global space which can be either a UBM model or an eigen-voice space. Adaptation methods which are included in this framework are MAP adaptation, maximum likelihood eigen decomposition (MLED) adaptation and MAP eigen-decomposition (MAPED) approaches (Kuhn et al., 2000; Huang et al., 2004).

Reference models selection: Not all the models in the generic model set are appropriate for employment in the proposed framework. The proposed reference model selection approach starts by computing KL_m distance between each two speaker models. After computing the distance between each two models, those two models which have the maximum distance from each other are selected. Then, the selected models are added to the set of reference models. In the next iteration, the model with the maximum distance from the reference models is selected and added to the reference set. The distance of a model from the reference set is denoted by the minimum distance between that model and each member of the set. This algorithm is repeated until the final number of reference models is reached. After selecting a predefined number of reference models, these models are organized in a tree structure as will be described in Section 12.1.

Silence removal: The indexing approach starts with silence removal. In silence removal, it is critical that we do not lose any speech data and minimize false rejection, even at the cost of false acceptance. In the proposed framework a voting based VAD algorithm is proposed (Moattar and Homayounpour, 2009). In this method, three different features are used per frame. The first feature is the widely used short-term energy. The second feature is spectral flatness measure (SFM) (Izmirli, 2000). Spectral Flatness is a measure of the noisiness of spectrum and is a good feature in Voiced/Unvoiced/Silence detection. Besides these two features, it was observed that the most dominant frequency component of the speech frame spectrum can be very useful in discriminating between speech and silence frames. The proposed algorithm starts with framing the audio signal. First 20 frames are used for threshold initialization. For each incoming speech frame, the above features are computed. The audio frame is marked as a speech frame, if more than one of the feature values exceed a pre-computed threshold. The pre-computed thresholds are determined on an arbitrary clean speech database (i.e. TIMIT corpus Garofolo et al., 1993). The complete VAD algorithm of the proposed framework is described in (Moattar and Homayounpour, 2009).

Feature extraction: The resulted silence removed stream is fed to the feature extraction module. The features used in this framework are 12-order MFCC vectors. To extract these features, a 30 ms Hamming window with 10 ms shift is used.

Speaker homogenous speech segmentation: For speaker segmentation two phases are designed. In this process, segmentation starts from higher granularity, and we aim to achieve the best possible recall. Therefore, the speech stream is segmented to short speech parts supposing that each segment has a unified content (i.e. speaker homogenous). To do so, some traditional speaker segmentation approaches are applied in parallel. Each of the segmentation methods, determine a set of locations as change point locations which may be different from the output of the other methods. Then, in the second step, the unreliable or less important change points are discarded using a voting and validation approach and the change point detection results are unified. Choosing the methods and the measures of homogenous segmentation depends on the definition of homogenous segments. In the proposed framework the homogenous segments are defined as follows:

1. Phonetic classes (***phone decoder based segmentation***): This approach segments the speech stream into phone level segments (Tranter et al., 2004) and, as mentioned in Section 6, over-segments the speech data. To avoid the dependency of the segmentation result to language dependent phone models, speech segmentation based on the language independent phonetic classes is used instead. For this purpose, the OGI multilingual database (Muthusamy et al., 1992) is used to train 3-state HMM models with 16 Gaussians in each state for each phone class. Seven phone classes are used which include, CLOS (closures (silence or background noise)), VOC (vowels), FRIC (fricatives), PRVS (pre-vocalic sonorant), POVS (post-vocalic sonorant), STOP (stops) and INVS

(inter-vocalic sonorant). This segmentation has the lowest level of granularity and Viterbi algorithm is applied for this purpose.

2. Utterances (***utterance based segmentation***): In this method the speech stream is divided into utterances. It is supposed that speaker change is occurred at the boundary of the utterances and not within them. However if this assumption is not true, the other approaches which are performed in parallel can enhance the output of the final system. There are two methods for segmentation in the proposed approach:
   (a) ***Pitch based segmentation***: In this approach the pitch contour is estimated. No smoothness or interpolation is performed on this contour. Then, the pitch contour is traversed and the places where discontinuities occur in the contour are marked as probable change points. The discontinuity may be caused by either an unvoiced segment in the speech stream or the differences in the pitch range of the succeeding speakers. To capture either of these events, a jump in the pitch contour with significant amplitude will be marked as change point.
   (b) ***Silence based segmentation***: In this method, as in traditional silence based approaches, the output of the VAD operation is used to determine the boundaries. The candidate speaker change point is supposed to be between each two speech segments.

3. ***Distance based speaker change detection***: Also, the classic distance based speaker segmentation methods are used in the proposed framework. Metrics such as Euclidean (EUC), Kullback-Leibler (KL), Mahalanobis (MAH), GLR and BIC are experimented in distance based segmentation. Also the SV novelty detector technique as described in (Desobry and Davy, 2003) is used in the proposed framework.

At each time slice, when the output of all of the above approaches is prepared, the change points are merged and verified using a voting based algorithm. In speaker ***change points refinement and verification***, the goal is to merge the output of the above methods. For this purpose, we consider a minimum distance between each two change points (with 500 ms duration) and also a predetermined number of votes for a valid change point in each time slice (20% of votes). The candidate point modification and verification algorithm is as follows:

Algorithm 1. Change points refinement and verification algorithm

1. The output of all of the segmentation methods is merged and sorted.
2. The resulted sequence is divided into the time slices with 500 ms duration.
3. In each time slice, if there exist enough votes gathered from all segmentation methods (20% of involved approaches), a candidate change point is considered. Otherwise no change point is considered in that time slice.
4. If a time slice have enough votes for having a change point, the average location of all change points in that slice, is considered as the location of the only change point in that slice.
5. The resulted change points are verified using modified BIC as applied in (Ajmera et al., 2004).

The above points are considered as the speaker boundaries and the speech utterance in between them is fed to the tracking module as the current segment to be classified.

Speaker tracking: At each step of the tracking, a homogenous speech segment is fed to the tracking procedure. The sequence of operations and decision makings in this phase, which are also illustrated in Fig. 10(b), are summarized in Algorithm 2:

Algorithm 2. Speaker tracking

1. Current speech segment is purified to exclude the frames which are likely to be silence or noise or the frames which are less informative (Anguera et al., 2006). First a GMM model is constructed for the current segment. Then the likelihood of each frame to the GMM model is computed and a number of speech frames with the highest likelihood are removed from the speech segment. The idea behind this approach is that the silence or noisy frames of a speech segment have the least variance and are clustered in the same component in the GMM model. Hence, when the likelihood is computed, the silence or noisy frames fall into the least variance components and receive the highest likelihood score.
2. The index tree of reference models is traversed to reach the nearest reference model to the current segment. To measure the similarity between current segment and reference models, likelihood ratio test is applied. When reaching one of the leaf nodes of the tree (i.e. a reference model), if a previously identified speaker model exists in that leaf, its model is updated with the current segment and the segment is labeled with a label corresponding to the nearest model. If the reached model is not an identified speaker, the tree structure is updated and a new speaker with an identical label is added to the set of currently found speakers.
3. The above steps are repeated until the data stream is exhausted.

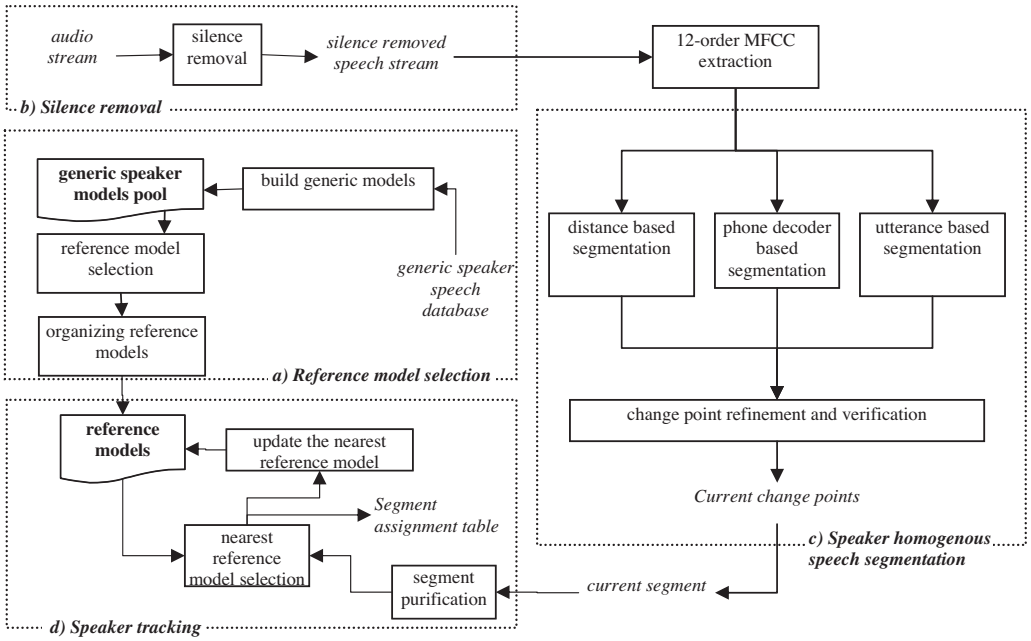

Fig. 10. Proposed online speaker indexing framework. (a) The generic speaker models are trained on a set of arbitrary speech dataset. The training algorithm can be substituted by adaptation of UBM model using the speaker specific data. (b) The audio stream is silence removed. (c) The speech stream is segmented to speaker homogenous parts. (d) In each time step, the input to the tracking algorithm is a homogenous speech segment which is surrounded by two detected change points. The output of tracking step is the labels assigned to speech utterances.

12.1. Organizing the speaker models

In the proposed approach, we aim to form a hierarchy of reference speakers by grouping N models bottom-up to obtain lower computational cost (i.e. $<O(N)$) and speeding up the tracking process. In this paper, we consider a binary tree as the indexing structure. In this structure, the two GMM speaker models $Model_1$ and $Model_2$ are represented by a single parent, $Model_{Merged}$, which also takes the form of a GMM. Let us denote the current segment in the tracking process by u_c. The cost of representing $Model_1$ and $Model_2$ by $Model_{Merged}$, can be expressed as follows:

$$\sum_{k=1,2} E[\ln p(u_c|Model_k)] - E[\ln p(u_c|Model_{merged})] \tag{25}$$

Let us define $Model_{Split}$ as the sum of the child models, in which the Gaussian components of both $Model_1$ and $Model_2$ are gathered under the same mixture. Supposing that the number of components in $Model_1$ and $Model_2$ are M_1 and M_2 respectively, the number of components in $Model_{Split}$ is $M_1 + M_2$. To maintain the summation of the weights of the components to 1, the weight coefficients of the component of $Model_{Split}$ are divided by 2. Assuming that both child models are equally probable, the loss of representing $Model_1$ and $Model_2$ by $Model_{Merged}$ can also be expressed by:

$$E[\ln p(u_c|Model_{Split})] - E[\ln p(u_c|Model_{Merged})] \tag{26}$$

The optimal model which minimizes the above cost is:

$$\hat{Model}_{Merged} = \arg\min_{sp} KL_m(Model_{Split}, Model_{Merged}) \tag{27}$$

where sp is the search space of all possible Gaussian components. In the above equation KL_m is the measure which is defined in Eq. (23). Eq. (27) means that the optimal parent model which can represent its child models best is the one that includes the components of $Model_{Split}$. $Model_{Split}$ has too many components to be computationally efficient. We should find the optimal mapping which maps $M_1 + M_2$ components of $Model_{Split}$ to M_{Merged} components of

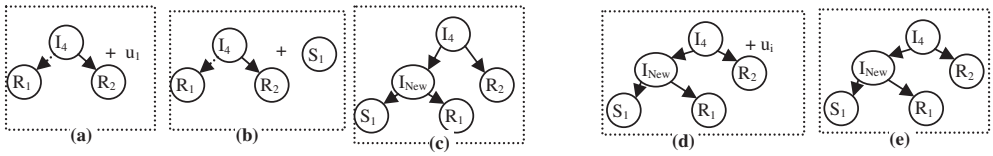

Fig. 11. The process of updating the tree structure with the entrance of a new utterance. (a) Current utterance, u_1, enters. u_1 is most similar to R_1. (b) R_1 is adapted with u_1. New S_1 model is generated. (c) Model insertion: since $R_1$is a reference model; R_1 and S_1 are reorganized as the child nodes of a new intermediate model, I_{New}. The parameters of I_{New} are computed using Algorithm 3. (d) A new utterance, u_i, is fed at time i. The most similar model to u_i is S_1. (e) S_1 is adapted with u_i and inserted back to the tree. No further reorganization is required.

$Model_{Merged}(M_{Merged} < M_1 + M_2)$ so that the KL_m divergence expressed in Eq. (27) is minimized.

To achieve this goal, an iterative algorithm similar to the classical k-means algorithm is applied. The proposed merging algorithm is summarized in Algorithm 3.

Algorithm 3. Merging algorithm for estimating $Model_{Merged}$ from $Model_1$ and $Model_2$

- Assign $M_1 + M_2$ Gaussian components to M_{Merged} clusters randomly.
- Repeat steps 1 and 2 until convergence.
  1. The parameters of the new mixtures are updated using the following equations:

$$\hat{w}_{j,Merged} = \sum_{i \in Set(j)} w_{i,Split} \tag{28}$$

$$\hat{\mu}_{j,Merged} = \sum_{i \in Set(j)} w_{i,Split}\mu_{i,Split}/\hat{w}_{j,Merged} \tag{29}$$

$$\hat{\Sigma}_{j,Merged} = \sum_{i \in set(j)} w_{i,Split}(\Sigma_{i,Split} + (\mu_{i,Split} - \hat{\mu}_{j,Merged}) \cdot (\mu_{i,Split} - \hat{\mu}_{j,Merged})^T)/\hat{w}_{j,Merged} \tag{30}$$

  where w_i, μ_i and Σ_i are the weight, mean vector and the variance vector of the ith Gaussian component respectively. Also $|\wedge|$ denotes the new estimation of the model parameters and $set(j)$ is the set of components which are assigned to the jth cluster in the current iteration.
  2. Reassign the "*Split*" components to the "*Merged*" components so that the KL divergence of this assignment is minimized.
- Update the parameters of the resulted Gaussian components.

Having Algorithm 3 for reorganizing the components of the child nodes in a single mixture of components as the parent model, the algorithm of organizing the reference models in an index tree structure is as follows:

Algorithm 4. Organizing the reference models in an index tree

Until only two models are remained (the root node is reached):
- Compute the KL_m distance between each two models using Eq. (23).
- Select two models with the least distance from each other.
- Use Algorithm 3 to merge the selected models in a single model.
- The resulted model is the parent of the previous models.

If the number of Gaussian components in each of the child models is M, for better presentation of the two speakers, the number of components in $Model_{Merged}$ would be greater than M. The number of Gaussian components in $Model_{Merged}$ should also be clearly smaller than $2M$ to ensure computational cost reduction. For simplicity and further cost reduction, we supposed that the number of components of the intermediate node models, $Model_{Merged}$, is equal to the number of components of the original reference models (i.e. M).

Since new speakers are incrementally added to the original set of reference models, the model indexing approach should be amenable to incremental processing and it can accommodate new speaker models. Since the process of model insertion is performed in the tracking time and one of the main requirements of tracking is low computational cost, the update process is only performed locally on the immediate parent. Fig. 11 explains the process of inserting a new adapted model into the tree. In this figure, the reference models are depicted with R and the intermediate models in the tree are depicted with I. Also, the new speaker model and the speech segments are depicted with S and u, respectively.

Another solution for constructing the intermediate nodes of the tree is experimented. This solution concerns the adaptation of the intermediate nodes using the data in their child nodes. In this solution the merging algorithm (Algorithm 3) is substituted with the adaptation of UBM model using the speech data of child nodes. This adaptation alternative should only be performed once and at the first formation of the index tree to maintain the tracking time.

12.2. Experiments

The main evaluation database is 2002 Rich Transcription Broadcast News (BN) and Conversational Telephone Speech (CTS). We have also generated a synthetic conversational dataset using the OGI multilingual database (Muthusamy et al., 1992) which contains speech utterances from 11 different languages. To generate this dataset, we used about 800 utterances of 80 speakers from 4 languages including English, Farsi, German and French and concatenated the utterances of different speakers in a random order. The synthetic utterances are generated with different number of speakers participating in each (i.e., 4, 8 and 40).

The first operation in the proposed framework is the construction of the generic models. Farsdat telephony and microphony speech datasets (Bijankhan, 2002) are applied for building generic models. Both of these datasets are in Farsi and are usually used for speech or speaker recognition over telephone and microphone, respectively. 50 generic speakers from the Farsdat telephony dataset and 200 generic speakers from the Farsdat microphony dataset are selected for CTS domain and BN domain, respectively. After silence removal from the speech utterances (Moattar and Homayounpour, 2009), a UBM model consisting of 64 components is trained using the speech data of generic speakers. To construct the model of each speaker, the UBM is adapted with the speech data of that speaker. To

adapt speaker models, 30 seconds of speech data of each speaker is used.

The experimental results are discussed separately. First set of experiments concern the speaker segmentation problem. In these experiments the aforementioned approaches for homogenous speaker segmentation and change points refinement are discussed. Then the proposed procedure for speaker tracking is evaluated. In the following experiments, evaluations are performed on CTS part of NIST 2002 Rich Transcription database, unless explicitly noted otherwise.

12.2.1. Homogenous speech segmentation

First experiments concern the speaker based speech segmentation task. For this purpose 6 different distance measures are experimented in a distance based framework. As discussed earlier, distance based segmentation approaches have various parameters (Siegler et al., 1997; Gauvain et al., 1998) from which the analysis window length (l), the overlap between two successive analysis windows (o), the shift (τ) between each two successive distance computation and the decision threshold (th) are the most important. Each of the mentioned parameters may significantly influence the segmentation accuracy. For example a long τ decreases the resolution of distance computation and may consequently decrease the change point recall. Also, the same phenomenon may happen when l or/and th increase. A similar influence may be caused by the penalty factor of the BIC metric, which as mentioned in Section 6.2.1.1, serves as a threshold. To exclude the BIC penalty factor, a similar approach as Ajmera et al. (2004) is employed in our experiments. To dynamically determine the proper th value, a formulation similar to Eq. (9) in Section 6.2.1.3 is applied. To obtain the highest number of change points (recall), the amplification coefficient, α, in Eq. (9) is set equal to 1. The other parameters are determined experimentally on all evaluation datasets. However, depending on the context to be segmented (i.e. BN or CTS), the proper values may be different, but these experiments try to catch up with a unified outcome.

Figs. 12 and 13 show the average RCL of different approaches versus their PRC. To get these results, l is changed between 500 ms to 1 s with 100 ms step size, o is changed between 50 ms to 350 ms which 50 ms step size and τ is set equal to 10ms to achieve the highest possible resolution and RCL. The experimented segmentation measures are divided to two groups just for better demonstration. Since the main goal of the homogenous speech segmentation is to achieve primary candidate change points with the least MDR, the priority is on the approaches which achieve the highest RCL even in the cost of low PRC.

Figs. 12 and 13 show that the best average RCL is achieved for Mahalanobis distance and then for the GLR. Also, the cost of this higher recall rate is not much for these two measures. The best resulted accuracies of the speaker segmentation process for different approaches versus different parameter settings are summarized in Table 2. Although the highest RCL is our favorite, but we have to maintain the PRC measure so that the resulted

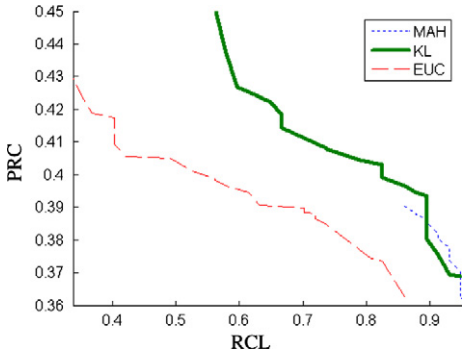

Fig. 12. Average accuracy of different distance metrics (i.e. MAH, KL, EUC) for speech segmentation on all evaluation databases in terms of RCL and PRC.

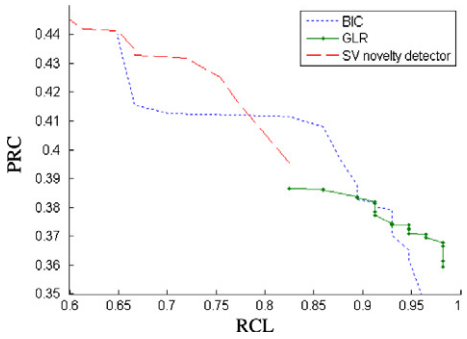

Fig. 13. Average accuracy of different distance metrics and approaches (i.e. BIC, GLR and SV novelty detector) for speech segmentation on all evaluation databases in terms of RCL and PRC.

Table 2
Average accuracy of distance based approaches on all evaluation databases in terms of RCL and PRC.

Distance measure	l (ms)	o (ms)	RCL (%)	PRC (%)
MAH	500	50	96.49	17.03
		250	98.25	16.14
	800	50	94.74	18.18
	900	150	96.49	19.03
KL	500	150	92.98	18.83
		350	95.74	16.88
	600	250	91.23	19.33
EUC	500	250	80.70	17.42
		350	85.96	16.28
	600	350	83.46	17.34
BIC	500	150	94.74	16.17
		350	92.98	21.54
	800	350	96.49	18.58
	900	50	92.98	20.78
GLR	500	50	98.25	18.36
	600	150	98.25	17.72
Sv-novelty detection	500	50	82.46	21.66

Table 3
Average accuracy of phone decoder based segmentation and pitch based segmentation on all evaluation datasets in terms of RCL, PRC.

Segmentation approach	RCL (%)	PRC (%)
Phone decoder based segmentation	100	11.95
Pitch based segmentation	78.95	25.57

speech segments are long enough for further processing and clustering. Also, it should be noted that in these experiments, the performance is achieved on all evaluation datasets including broadcast news and CTS with intrinsically different length of utterances.

The resulted RCL in Table 2 is considerably higher than the state-of-the-art segmentation accuracies (Kotti et al., 2008). On the other hand, the precision of the segmentation is very low. Actually, in average, there are about 5 (i.e. 4.46) false alarms for each truly detected change points. Actually, it means that the segmentation algorithm over-segments the speech signal, but this is the cost we should pay to get the highest recall rate. However, these false alarms can be relatively removed by further change point refinement and also in the tracking algorithm. The highest RCLs are analogous to the shortest window length (l), but there is not a definite trade-off between parameter o and the segmentation accuracy. Also, it is shown that the Euclidean distance is the worst approach and the sv-novelty detection approach is the most vulnerable method against short analysis window length.

Table 2 shows that the most robust and accurate distance based approaches are Mahalanobis and GLR. However, GLR distance computation is computationally inefficient, and this approach is less desirable than Mahalanobis when the computational complexity is considered. Therefore, the selected distance measure in the proposed homogenous segmentation strategy is Mahalanobis. The reader may consider that the purpose of these experiments is to select an appropriate approach that satisfies high recall rate and moderate precision and we do not intend to compare the total performance of speaker segmentation approaches. Therefore, we have dissected out F measure from our evaluations.

The other two homogenous speech segmentation approaches namely phone decoder based segmentation and pitch based segmentation are also evaluated. Table 3 shows the results of these experiments. These segmentation approaches do not have any specific parameters and the results are independent from parameter tunings.

The 100% RCL of phone decoder based approach and its very low PRC is reasonable; because this approach segments the speech stream to the smallest homogenous segments possible and hence the resulted candidate points are very close to the actual change points. Also, as expected, the RCL of the pitch based segmentation approach is the lowest, while its PRC is relatively high. The reason is similar to the reason of low RCL of VAD based approaches, in which there is not a complete analogy between each change point and silence region.

Table 4
Average accuracy of the homogenous speaker segmentation procedure on all evaluation databases after change point refinement and verification in terms of RCL, PRC.

Change point refinement	RCL (%)	PRC (%)
Before BIC validation (only voting is applied)	96.49	23.15
After BIC validation	91.23	35.55

12.3. Change point refinement and validation

To benefit the advantages of all the above approaches, and increase the PRC measure as much as possible the classifier to be trained to the tracking phase while maintaining recall rate, the best of the above methods are combined in a voting and refinement framework introduced earlier, to have a unified output.

In these experiments, the best distance based approach, which is Mahalanobis based segmentation, is applied in parallel with phone decoder segmentation and pitch based segmentation while the output of the VAD algorithm (Moattar and Homayounpour, 2009) is also used. Algorithm 1 is applied to unify the outputs of these approaches. Although the mentioned algorithm seems to be so simple and elementary, but it is relatively fast and can remove some of the wrongly detected change points while maintaining the recall to an acceptable rate. Table 4 shows the results of these experiments.

Whether using the BIC validation or not, the number of false alarms per detected change points, decrease to 1.81 and 3.31, respectively. This means that the average length of the speaker utterances is increased which is so helpful in tracking procedure. The output of this phase is actually a set of candidate change points which will be removed or refined in the tracking phase. The speech stream between each two detected change points is fed to the tracking system to be assigned to a proper speaker. But, before we begin our discussion on the tracking procedure, we may need to discuss two main questions. First, what is the most appropriate number of reference speaker models, while the actual number of speakers in conversations is unknown?

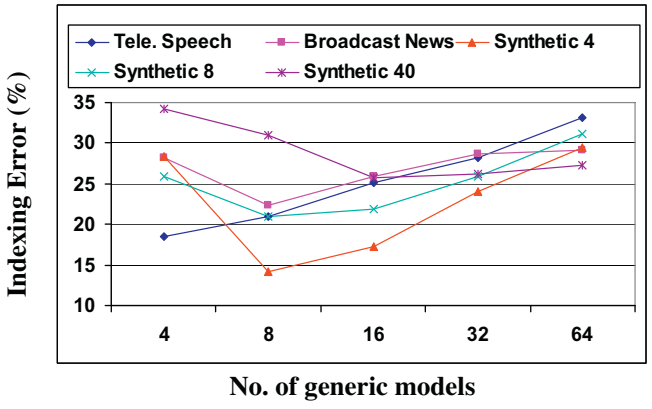

Fig. 14. Indexing error of the proposed framework on different datasets for various number of reference models. Synthetic k means that the number of speakers in the synthetic conversations is k.

Second, what is the best approach to build the generic models? Does different approaches, for generic model training and adaptation influences the indexing results? In other word, does the nature of the primary adaptation space (i.e. a universal space for speakers (GMM-UBM) or eigen-voice space) influences the discriminative capability of the models and enhances the results? These two questions are discussed in the next subsections.

12.3.1. Number of reference models

It is not rational to set the optimal number of reference models using the test data because it is suggested that there is no information about the actual nature of the input conversation. However, it will be useful to monitor the indexing accuracy on different test sets for various reference model numbers. Fig. 14 illustrates the indexing error of the proposed approach on three different datasets versus the number of reference speaker models. We considered different sizes for reference models set (i.e. 4, 8, 16, 32 and 64 reference models). In the following experiments, the indexing error is measured as Eq. (24) in Section 11. Also, in the experiments of the current and the following subsections the set of reference speaker models are exhaustively searched and the index structure is relinquished.

Fig. 14 shows that the average indexing error of the proposed approach increases by the increase in the number of reference models. Only when the number of speakers in the conversation is high, as in 40 speaker conversations, increasing the number of reference speaker models may improve the results. We have to consider that using a fixed-size set of reference models for various applications is a double-edged sword. Few reference speakers may cause misclassification to wrong speakers. Also, a very large reference set may cause the utterances of a single speaker to be divided between two or more classes. Hence, a moderate number of speakers seems to be better. From the above experiments, we can conclude that 8 to 16 reference speakers is optimum. In the future experiments, 8 reference models are used, although 16 reference speakers could also be evaluated

12.3.2. Training generic models

In experiments of this section, we evaluated the influence of the generic speaker model training methods on the indexing error. For this purpose, the effect of eigen-voice adaptation methods (i.e. MLED and MAPED) Kuhn et al., 2000; Huang et al., 2004 is compared with MAP adapted models.

Table 5
Indexing error rate on CTS data using different adaptation method to build the generic models.

Search method	Reference model adaptation approach	Indexing error (%)
Exhaustive search	MAP	16.44
	MLED	10.33
	MAPED	9.00

The eigen-voice adaptation approach has two steps which are eigen-voice construction and coefficient estimation. The second phase is usually known as adaptation phase. In the first phase, a set of speaker models from G speakers is trained using conventional EM algorithm. For each model, the mean vectors of the Gaussian components are concatenated to form a supervector. Suppose that the dimension of the feature space is d and the number of components in the GMM model is denoted by M. Hence, the dimension of the resulted supervector is $D = d^{*}M$. The $D^{*}D$ covariance matrix is calculated from G supervectors. Then, eigen decomposition (i.e. principle component analysis (PCA)) is applied on this covariance matrix. The first K more dominant eigenvectors are selected to form a K-dimensional eigen-space. The selected eigenvectors are called eigen-voice in literature. In the eigen-voice space, the mean supervector of a speaker is represented with a linear combination of K eigen-voices. The aim of eigen-voice adaptation approaches is to find these combination coefficients. For this purpose, various criteria can be applied from which ML criterion (i.e. MLED) and MAP criterion (i.e. MAPED) are used in our experiments.

Eigen-voice adaptation approaches are said to be robust against data insufficiency (Huang et al., 2004). In the proposed framework, we apply these approaches for adapting primary generic speaker models. We suppose that using these approaches, the spatial locations of the generic speakers will also be embedded in the corresponding adapted models. This will help to select the most spatially different speakers in the selection algorithm and span the speaker space more efficiently. The results of these experiments are summarized in Table 5.

In these experiments, the set of generic speaker models is used to build the eigen-space. As mentioned previously, each supervector is constructed by concatenating the mean vectors of the corresponding GMM model. Since GMM models have 64 mixtures and the dimension of the feature vectors is 12, the dimension of supervectors, D, is 768 (64*12). After PCA eigen decomposition, the first 33 eigen-voices are used as the bases of the eigenspace.

Table 5 shows that the MAPED and MLED approaches are better than the MAP adaptation. As mentioned before, eigen-voice space represents the directions of speaker variations. When speaker models are constructed using the

Table 6
Indexing error of different search strategies on CTS data. The experiments are performed for different reference model adaptation approaches and intermediate node generation methods.

Search method	Reference model adaptation approach	Intermediate node training and update	Indexing error (%)
Tree search	MAP	Algorithm 3	19.46
	MLED		10.51
	MAPED	Algorithm 3	9.36
		MAP	11.42
		MLED	9.00
		MAPED	7.51

linear combination of eigen-voices (i.e. adaptation), they rest in the orientation of the speaker variations. Therefore, the adapted speaker models embed the structure of the speaker space and span the speaker space better. Clearly, when the speaker space is spanned better and the generic models are more diverse, the tracking accuracy will be improved.

12.3.3. Index structure

In the final experiments, we evaluated the effect of the index tree of the reference models on the accuracy of the indexing. The results of applying this structure in the indexing procedure are summarized in Table 6. The following results are extracted on the CTS test set.

These experiments show that the first tree construction approach which uses Algorithm 3 increases the indexing error rate compared to the exhaustive search (indicated in Table 5). It is due to the fact that the model which is resulted from Algorithm 3 through the merging scheme is weak and can not represent its child good enough. On the other hand, the adaptation approaches for constructing the intermediate nodes of the tree decreases the indexing error. This improvement is the highest when both reference models and the nodes of the tree are trained using the MAPED adaptation. The reason behind this better accuracy is that the adaptation approach constructs better parent model and hence the cost function of representing two models with their parent model decreases to the minimum. Since the index tree formation is an offline procedure, the runtime of the tracking operation will be indifferent using either of these methods to construct the tree.

12.3.4. Final evaluation

The indexing error rate on all three evaluation datasets using the proposed framework is illustrated in Table 7. The indexing error on the synthetic dataset is averaged on all three subsets.

As illustrated in Table 7, the indexing error is almost the same for different application domains which shows the robustness of the indexing procedure. Only the error rate on the synthetic dataset is relatively higher which is due to the diversity of speakers and languages in these conversations.

Table 8 tries to illustrate the effectiveness of the proposed index tree structure from the processing time point of view. These processing times are achieved on an AMD Athlon Dual-Core processor with 2.7 GHz clock rate and 3 GB of RAM for each second of input speech. These results only include the segment assignment and tree update operations and the process time of fixed tasks such as silence removal, feature extraction, change detection and frame purification are excluded from these experiments.

These results show that the proposed tree structure does not improve the process time of the tracking task much, when the number of speakers in the speech data is low (i.e. telephone conversation or 4 speaker conversations). The main improvement arises when the number of speakers in the speech documents increases. This enhancement is clearly seen for broadcast news and 40 speaker synthetic conversations. Also, we can conclude that the assignment time of the index tree structure is almost linear while the process time of exhaustive search increases exponentially by the number of speakers.

To show the applicability of the proposed approach in online applications, we have to demonstrate that the run time of the proposed framework satisfies real-time requirements and its real-time factor (xRT) is lower than 1. Table 9 shows the average real-time factor of each module of the proposed framework.

Since the average xRT for the proposed speaker tracking (using index tree structure) on different datasets is 0.338 (Table 8), then the average xRT of the whole framework will be 0.434. This shows that processing time of the proposed framework meets the real-time requirements.

Table 7
Indexing error rate on all three datasets using the proposed framework for both exhaustive search and index tree search.

Evaluation database	Indexing error (%)	
	Exhaustive search (without index tree)	Index tree search
Telephone conversation	9.00	7.51
Broadcast news	9.07	6.36
Synthetic dataset	10.61	9.34

Table 8
Average process time of assigning each second of speech to its corresponding speaker (milliseconds).

	Tele. speech	Synthetic 4	Synthetic 8	Broadcast news	Synthetic 40
Linear search	380	409.7	421.6	455.1	523.1
Tree structure	330	330.6	335.5	345.2	351.3
Improvement (%)	13.15	19.30	20.42	24.14	32.84

Table 9
Average real-time factor (xRT) of different modules of the proposed framework.

	Silence removal	Feature extraction	Change detection module		
			Distance based	Decoder based	Utterance based
xRT	0.014	0.020	0.031	0.018	0.013

13. Most important future research direction

In this section, we review those areas of diarization technology which are still open and can improve diarization performance. We first discuss detecting the overlapping speech as the main bottleneck of many diarization systems. Then, the use of prosodic information for promising speaker diarization results is discussed. Also, multimodal speaker diarization and computational complexity reduc-

tion of the approaches are some of the other main directions of future works in this area.

13.1. Overlap detection

Some studies have been reported about the effects of overlap in meetings, but works on systems for identifying overlapped speech and investigating its effects in speaker diarization are rare in the literature. As overlapped speech is now a major obstacle in improving the performance of speaker diarization systems, efforts in overlap detection will be of increasing interest and importance.

A fundamental limitation of most current speaker diarization systems is that only one speaker is assigned to each segment. However, the presence of overlapped speech is common in multiparty meetings and presents a significant challenge to automatic systems. Specifically, in regions where more than one speaker is active, missed speech errors will be incurred. Segments which contain speech from more than a single speaker should not be assigned to any individual speaker cluster nor included in any individual speaker model. Doing so adversely affects the purity of speaker models, which ultimately reduces diarization performance.

Approaches to overlap detection were thoroughly assessed in (Shriberg et al., 2001; Çetin and Shriberg, 2006). Only a small number of systems actually detect overlapping speech well enough to improve error rates (Boakye et al., 2008; Trueba-Hornero, 2008; Boakye, 2008). The main approach for overlap detection consists of a three state HMM-GMM system (non-speech, non-overlapped speech, and overlapped speech), and the best feature combination is MFCC and modulation spectrogram features (Kingsbury et al., 1998. In Trueba-Hornero (2008) a real overlap detection system was developed, as well as a better heuristic that computed posterior probabilities from diarization to post process the output and include a second speaker on overlap regions. The main bottleneck of the achieved performance gain is mainly due to errors in overlap detection, and more work on enhancing its precision and recall is reported in (Boakye et al., 2008; Boakye, 2008).

Other suggested methods for detecting overlapping speech segments include the use of support vector regression (SVR) (Kotti et al., 2008); and the extraction of acoustic features to be used with a GMM, which assumes that each participant has an individual microphone, and requires the classifier to be trained for each combination of features. An alternative approach to detecting the presence of overlapping speech is to estimate the number of present sources. Any segment containing more than one source signal can then be classified as overlapping speech. In some cases the second present sources may not actually be a speech source and might instead be due to laughter or coughing. However, as these situations also lead to degradation of speech recognition techniques it is likewise important to identify such sections of the recording. Classical source number determination techniques are based on eigen-decomposition of the observed spatial correlation matrix. Under ideal conditions (i.e. high SNR), the eigenvalues corresponding to the noise subspace are equal and the number of present sources is easily determined as the number of non-equal eigenvalues. The most well known source number determination techniques are the Akaike information criterion (AIC) (Zhu et al., 2008) and Rissanen's minimum description length (MDL) (Jothilakshmi et al., 2009). However, due to the difficult operational conditions encountered in meeting recordings they are no longer accurate, as even at high Signal to Noise Ratio (SNR), they continuously over-estimate the number of present sources (Rich Transcription Meeting Recognition Evaluation Plan, 2009).

Detection of overlapping speech segments is a difficult problem due to both the nature of the speech and the environmental effects such as background noise, echo and reverberation (Mirghafori and Wooters, 2006). Moreover, any suitable algorithm should be computationally simple, in order to allow for real-time implementation (Anguera, 2006), and the amount of training data required should be minimal.

13.2. Use of prosodic features

The use of prosodic features for both speaker detection and diarization is emerging as a reaction to the inconsistency of using MFCC features (Wölfel et al., 2009; Friedland et al., 2009). In (Friedland et al., 2009) the authors present a systematic investigation of the speaker discriminability of 70 long-term features, most of them prosodic features. They provide evidence that despite the dominance of short-term cepstral features in speaker recognition, a number of long-term features can provide significant information for speaker discrimination. The consideration of patterns derived from larger segments of speech can reveal individual characteristics of the speakers' voices which cannot be captured using a short-term, frame-based cepstral analysis (Shriberg, 2007). Therefore, the authors of Friedland et al. (2009) use Fisher LDA as a ranking methodology and sort the 70 prosodic and long-term features by speaker discriminability. The combination of the top-ten ranked prosodic and long-term features combined with regular MFCCs leads to improvement in terms of DER.

13.3. Audiovisual diarization

Audiovisual diarization, which is discussed in Section 10, is a novel emerging area in speaker diarization field. The state of the art approaches can produce high accuracy speaker diarization outputs. However despite of the works done in this field, there are still areas which are not covered properly. In previous audiovisual approaches, there has been no effort to combine multimodal diarization and speaker detection algorithms with other modalities such as speaker location for speaker diarization. Also actual synchrony between the audio and video is rarely used

and the audio and video streams are treated as independent sources of information. Another concern of multimodal diarization is that multiple cameras and microphones are not available in most recordings. This is the main limitation of applying the state of the art multimodal diarization systems and approaches in real applications.

13.4. Diarization speed-up

A main concern in speaker diarization is the time complexity of the different approaches involved. The time complexity of speaker diarization approaches not only concerns the main steps of speaker diarization such as speaker segmentation and tracking, but also it is very important to have the least computations in other parts of the diarization framework. This computation reduction is necessary especially for online and real-time diarization tasks. Diarization speeding-up would start from the first step which is voice activity detection. Robust and real-time approaches for this purpose are necessary. Two examples for these approaches are those that are proposed in (Moattar and Homayounpour, 2009 and Yoo and Yook, 2009). Computational complexity in feature extraction phase may be reduced by using lower order features (Kotti et al., 2008) or efficient dimension reduction approaches (Kinnunen et al., 2008). Also, in the core speaker diarization framework approaches such as uniform segmentation or fast adaptation and update approaches can be applied. Approaches such as the index tree model search in tracking or fast parallel speaker segmentation which are applied in the proposed framework and are discussed and evaluated in Section 12, can be some possible choices. From the other previous works for this purpose, the approaches applied in (Rougui et al., 2006; Anguera and Bonastre, 2011; Huang et al., 2007) can be mentioned.

14. Conclusions

This paper presents a complete and detailed review on the approaches proposed for different aspects of speaker based indexing and diarization in the literature. The studied approaches are categorized based on their basic idea and contributions. This review paper provides the basic materials and ideas for researchers. Also, this paper proposes an indexing system which gathers the most widely used indexing approaches and some novel ideas under a unified framework for application domain independent online speaker indexing.

From the mentioned discussions, we can conclude that there was a great interest in the field of speaker diarization including measures and algorithms and also its applicability. Therefore, the diarization systems have opened their way through speech analysis and retrieval applications such as speech recognition (Stolcke et al., 2010). Today, there are many open-source toolboxes freely available which are dedicated to one or more aspects of speaker indexing and implement most of the previous methods and algorithms including feature extraction, silence removal, speaker segmentation and speaker tracking and clustering. Research on speaker diarization has been developed in many domains, from phone calls conversations within the speaker recognition evaluations, to broadcast news and meeting recordings. Furthermore, it has been used in many applications such as a front-end for speaker and speech recognition, as a meta-data extraction tool to aid navigation in broadcast TV, lecture recordings, meetings, and video conferences. In addition, the availability of other modalities, such as video, has started to inspire multimodal diarization systems, thus merging the audio and visual domains. Therefore, new efforts are mostly biased towards real word systems which provide the users with applicable metadata and are also usable in complement with other speech processing and understanding systems and applications such as movie analysis and speech transcription.

A more systematic study should be carried out to identify effective features for the task. Thus far, traditional acoustic features, such as MFCC and LPCC which capture short-term statistics of speech frames, are being used. Short-term cepstral systems generally perform well because they reflect information about the speakers' physiology and do not rely on the phonetic content. However, long range information has the potential of increased robustness to channel variation, since lexical usage or temporal patterns do not change with the change of acoustic conditions. Utilizing prosodic information can be another horizon for future works. Systems currently do not exploit such features as pitch trajectories and other super-segmental features to aid in segmentation or clustering. Prosodic features capture speaker specific information. Thus, they probably are good candidates for combination with short-term features in speaker diarization systems. Given multiple feature streams, it is also noteworthy to examine more effective frameworks to combine these individual feature streams.

Also, more robust approaches against data shortage in online applications and the approaches for determining the actual number of speakers are still open problems in speaker diarization. Also, speaker segmentation and clustering algorithms are mostly proposed for performing the tasks in clean conversations, and their efficiency may decrease in adverse conditions. However, clean conditions are the ideal cases and there are many conversations which take place in noisy environments. Some examples for such environments are broadcast news reports in crowded places and long distance telephone conversations. Also, larger datasets need to be provided in order to achieve more meaningful results and for systems to be more robust to unseen variations. Of course, with increasing dataset sizes, systems will have to become computationally more efficient in order to process such data in reasonable time. Also, there is still more that can be done towards eliminating as many tuning parameters as possible, allowing algorithms to obtain such parameters only from the data.

Still, the biggest challenge is to handle overlapping speech, that needs to be assigned to multiple speakers. Another area where the research should be directed is the creation of strong links between ASR transcription output and diarization. Although the use of diarization algorithms to help ASR systems in model adaptation is well established, the use of ASR for diarization has been briefly studied.

In the current state-of-the-art diarization systems, it is assumed that no prior information about the number of speakers is provided. It would be interesting to explore particular areas of application where the number of speakers is known. This particular information may change the way the speaker diarization algorithms are designed and some techniques of speaker identification may be applicable.

Overall, the future of the field seems even broader and brighter than the present, as more and more people acknowledge the usefulness of audio methods for many tasks that have traditionally been thought to be exclusively solvable in the visual domain. Speaker diarization is one of the fundamental problems underlying virtually any task that involves acoustics in the presence of more than one person.

Acknowledgements

The authors would like to thank Iran Telecommunication Research Center (ITRC) for supporting this work under contract No. T/500/14939.

References

Ajmera, J., Wooters, C., 2003. A robust speaker clustering algorithm, In: Proc. of IEEE ASRU Workshop, St. Thomas, USA, 2003, pp. 411–416.

Ajmera, J., Bourlard, H., Lapidot, I., McCowan, I., 2002. Unknown multiple speaker clustering using HMM, In: Proc. of ICSLP, USA, pp. 573–576.

Ajmera, J., Bourlard, H., Lapidot, I., 2002. Improved unknown-multiple speaker clustering using HMM, IDIAP Research Report, Martigny, Switzerland.

Ajmera, J., Lathoud, G., Mc-Cowan, I., 2004. Clustering and segmenting speakers and their locations in meetings, In: Proc. of ICASSP 1, pp. 605–608.

Ajmera, J., McCowan, I., Bourlard, H., 2004. Robust speaker change detection. IEEE Signal Process. Lett. 11 (8), 649–651.

Akita, Y., Kawahara, T., 2003. Unsupervised speaker indexing using anchor models and automatic transcription of discussions, In: Proc. of EUROSPEECH, pp. 2985–2988.

Alabiso, J., MacIntyre, R., Graff, D., 1997. English Broadcast News Transcripts (HUB4), Linguistic Data Consortium, Philadelphia, 1998.

AMI, Augmented Multi-party Interaction, Available: <http://www.amiproject.org>.

Anguera, X., 2006. Robust speaker diarization for meetings, Ph.D. Thesis, Universitat Politecnica de Catalonia.

Anguera, X., Bonastre, J.-F., 2011. Fast speaker diarization based on binary keys, In: Proc. of ICASSP.

Anguera, X., Hernando, J., 2004. XBIC: nueva medida para segmentacion de locutor hacia el indexado automatico de la senal de voz, in III Jornadas en Tecnologia del Habla, Valencia, Spain.

Anguera, X., Wooters, C., Peskin, B., Aguiló, M., 2005. Robust speaker segmentation for meetings: The ICSI-SRI Spring 2005 diarization system, In: Proc. of Machine Learning for Multimodal Interaction Workshop (MLMI), Edinburgh, UK, pp. 402–414.

Anguera, X., Wooters, C., Peskin, B., Aguilo, M., 2006. Robust speaker segmentation for meetings: the ICSI-SRI spring 2005 diarization system. Proc. of Machine Learning for Multimodal Interaction (MLMI). In: Lecture Notes in Computer Science. Springer-Verlag, Berlin, pp. 402–414.

Anguera, X., Aguilo, M., Wooters, C., Nadeu, C., Hernando, J., 2006. Hybrid speech/non-speech detector applied to speaker diarization of meetings, In: Proc. of Odyssey, pp. 1–6.

Anguera, X., Wooters, C., Hernando, J., 2006. Frame purification for cluster comparison in speaker diarization, In: Proc. of Second Internat. Workshop on Multimodal User Authentication.

Anguera, X., Wooters, C., Pardo, J., 2006. Robust speaker diarization for meetings: ICSI RT06s evaluation system, In: Proc. of Ninth Internat. Conf. on Spoken Language Processing, ISCA.

Anguera, X., Wooters, C., Pardo, J., 2007. Robust speaker diarization for meetings: ICSI RT06s evaluation system. Proc. of Machine Learning for Multimodal Interaction (MLMI). In: Lecture Notes in Computer Science, vol. 4299. Springer-Verlag, Berlin.

Anguera, X., BeamformIt: the fast and robust acoustic beamformer, <http://www.xavieranguera.com/beamformit/>.

Antolin, A.G., Anguera, X., Wooters, C., 2007. Speaker diarization for multiple-distant-microphone meetings using several sources of information. IEEE Trans. Comput. 56 (9), 1212–1224.

Arias, J.A., Pinquier, J., Obrecht, R.A., 2005. Evaluation of classification techniques for audio indexing, In: Proc. of the 13th European Signal Processing Conf., Antalya, Turkey.

Bakis, R., Chen, S., Gopalakrishnan, P., Gopinath, R., 1997. Transcription of broadcast news shows with the IBM large vocabulary speech recognition system, In: Proc. of Speech Recognition Workshop, pp. 67–72.

Barras, C., Gauvain, J.-L., 2003. Feature and score normalization for speaker verification of cellular data, In: Proc. of ICASSP II, China, pp. 49–52.

Barras, C., Zhu, X., Meignier, S., Gauvain, J.-L., 2004. Improving speaker diarization, In: Proc. of Fall Rich Transcription Workshop (RT-04), USA.

Barras, C., Zhu, X., Meignier, S., Gauvain, J.L., 2006. Multistage speaker diarization of broadcast news. IEEE Trans. Audio Speech Language Process. 14 (5), 1505–1512.

Basseville, M., Nikiforov, I., 1993. Detection of Abrupt Changes: Theory and Application. Prentice-Hall.

Ben-Harush, O., Lapidot, I., Guterman, H., 2008. Weighted segmental k-means initialization for SOM-based speaker clustering, In: Proc. of INTERSPEECH, pp. 24–27.

Biatov, K., Köhler, J., 2006. Improvement speaker clustering using global similarity features, In: Proc. of INTERSPEECH, USA, pp. 2082–2085.

Bijankhan, M., 2002. Great Farsdat database, Technical report, Iran Research center on Intelligent Signal Processing.

Bimbot, F., Mathan, L., 1993. Text-free speaker recognition using an arithmetic-harmonic sphericity measure, In: Proc. of Eurospeech, Germany, pp. 169–172.

Boakye, K., 2008. Audio Segmentation for Meetings Speech Processing, Ph.D. Dissertation, University of California at Berkeley.

Boakye, K., Trueba-Hornero, B., Vinyals, O., Friedland, G., 2008. Overlapped speech detection for improved speaker diarization in multiparty meetings, In: Proc. of ICASSP, pp. 4353–4356.

Boehm, C., Pernkopf, F., 2009. Effective metric-based speaker segmentation in the frequency domain, In: Proc. of ICASSP, pp. 4081–4084.

Bozonnet, S., Evans, N.W.D., Fredouille, C., 2010. The LIA-Eurecom Rt'09 speaker diarization system: enhancements in speaker modelling and cluster purification, In: Proc. of ICASSP, pp. 4958–4961.

Brandstein, M., Griebel, S., 2001. Explicit Speech Modeling for Microphone Array Applications. Springer, Chapter 7.

Burger, S., Maclaren, V., Yu, H., 2002. The ISL meeting corpus: The impact of meeting type on speech style, In: Proc. of ICSLP, Denver, USA, pp. 301–304.

Campbell, J.P., 1997. Speaker recognition: a tutorial, In: Proc. of the IEEE 1.85 9, pp. 1437–1462.

Campbell, N., Suzuki, N., 2006. Working with very sparse data to detect speaker and listener participation in a meetings corpus, In: Proc. of Workshop Programme.

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., Vair, C., 2008. Stream-based speaker segmentation using speaker factors and eigenvoices, In: Proc. of ICASSP, pp. 4133–4136.

Çetin, O., Shriberg, E., 2006. Speaker overlaps and ASR errors in meetings: effects before, during, and after the overlap. Proc. of ICASSP, 357–360.

Cettolo, M., Vescovi, M., 2003. Efficient audio segmentation algorithms based on the BIC, In: Proc. of ICASSP, pp. 537–540.

Cettolo, M., Vescovi, M., Rizzi, R., 2005. Evaluation of BIC-based algorithms for audio segmentation, Comput. Speech Language 19, pp. 1004–1013.

Chen, S.S., Gopalakrishnan, P.S., 1998. Clustering via the Bayesian information criterion with applications in speech recognition, In: Proc. of ICASSP 2, USA, pp. 645–648.

Chen, T., Rao, R., 1996. Cross-modal prediction in audio-visual communication, In: Proc. of ICASSP 4, pp. 2056–2059.

Chen, S.S., Gales, M.J.F., Gopinath, R.A., Kanvesky, D., Olsen, P., 2002. Automatic transcription of broadcast news. Speech Commun. 37, 69–87.

Cheng, S., Wang, H., 2004. Metric SEQDAC: a hybrid approach for audio segmentation, In: Proc. of the 8th International Conference on Spoken Language Processing, Jeju, Korea, pp. 1617–1620.

CHIL, Computers in the Human Interaction Loop, Available: <http://chil.server.de/>.

Chu, S.M. et al., 2008. Recent advances in the IBM GALE mandarin transcription system, In: Proc. of ICASSP, pp. 4329–4332.

Chu, S.M., Tang, H., and Huang, T.S., 2009. Fishervoice and semi-supervised speaker clustering, In: Proc. of ICASSP, pp. 4089–4092.

Cohen, I., Berdugo, B., 2002. Speech enhancement based on a microphone array and logspectral amplitude estimation, In: Proc. of 22nd Convention of Electrical and Electronics Engineers in Israel.

Collet, M., Charlet, D., Bimbot, F., 2003. A correlation metric for speaker tracking using anchor models, In: Proc. of ICASSP 1, Hong Kong, pp. 713–716.

Cox, H., Zeskind, R., Kooij, I., supergain, Practical, 1986. Practical supergain. IEEE Trans. Acoust. Speech Signal Process. 34 (3), 393–397.

Cox, H., Zeskind, R., Owen, M., 1987. Robust adaptive beamforming. IEEE Trans. Acoust. Speech Signal Process. 35 (10), 1365–1376.

Delacourt, P., Wellekens, C.J., 2000. DISTBIC: a speaker based segmentation for audio indexing. Speech Commun. 32 (1–2), 111–127.

Delacourt, P., Kryze, D., Wellekens, C.J., 1999. Detection of speaker changes in an audio document. Proc. Eurospeech 3, 1195–1198.

Delphine, C., 2010. Model-free anchor speaker turn detection for automatic chapter generation in broadcast news, In: Proc. of ICASSP, pp. 4966–4969.

The Department of Speech, Music and Hearing of the Royal Institute of Technology (KTH) at Stockholm, <http://www.speech.kth.se>.

Deshayes, J., Picard, D., 1986. Online Statistical Analysis of Change-point Models Using Non-parametric and Likelihood Methods. Springer-Verlag.

Desobry, F., Davy, M., 2003. Support vector-based online detection of abrupt changes, In: Proc. of ICASSP 5, pp. 872–875.

El-Khoury, E., Sénac, C., Pinquier, J., 2009. Improved speaker diarization system for meetings, In: Proc. of ICASSP, pp. 4097–4100.

Ellis, D.P.W., Liu, J.C., 2004. Speaker turn segmentation based on between-channel differences, In: Proc. of NIST Meeting Recognition Workshop at ICASSP 2004.

Evans, N.W.D., Fredouille, C., Bonastre, J.F., 2009. Speaker diarization using unsupervised discriminant analysis of inter-channel delay features, In: Proc. of ICASSP, pp. 4061–4064. .

Fernandez, D., Otero, P.L., Mateo, C.G., 2009. An adaptive threshold computation for unsupervised speaker segmentation, In: Proc. of Interspeech, Brighton, UK, pp. 849–843.

Fischer, S., Kammeyer, K.D., 1997. Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environments, In: Proc. of ICASSP.

Fiscus J.G. et al., 2004. Results of the fall 2004 STT and MDE evaluation, In: Proc. of Fall 2004 Rich Transcription Workshop (RT-04), Palisades, NY.

Fiscus, J.G., Radde, N., Garofolo, J.S., Le, A., Ajot, J., Laprun, C., 2005. The rich transcription 2005 spring meeting recognition evaluation, In: Proc. of Machine Learning for Multimodal Interaction Workshop (MLMI), Edinburgh, UK, pp. 369–389.

Fisher, J.W., Darrell, T., 2004. Speaker association with signal-level audiovisual fusion. IEEE Trans. Multimedia 6 (3), 406–413.

Fisher, J.W., Darrell, T., Freeman, W.T., Viola, P.A., 2000. Learning joint statistical models for audio-visual fusion and segregation, In: Proc. of NIPS, pp. 772–778.

Flanagan, J. et al., 1994. Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Amer. 78, 1508–1518.

Fredouille, C., Senay, G., 2006. Technical improvements of the E-HMM based speaker diarization system for meeting records, In: Proc. of MLMI'06, Washington, USA.

Fredouille, C., Moraru, D., Meignier, S., Besacier L., Bonastre, J.F., 2004. The NIST 2004 spring rich transcription evaluation: Two-axis merging strategy in the context of multiple distant microphone based meeting speaker segmentation, In: Proc. of NIST 2004 Spring Rich Transcription Evaluation Workshop, Montreal, Canada.

Friedland, G., Vinyals, O., Huang, Y., Muller, C., 2009. Prosodic and other long-term features for speaker diarization. IEEE TASLP 17 (5), 985–993.

Friedland, A.G., Vinyals, B.O., Huang, C.Y., Muller, D.C., 2009. Fusing short term and long term features for improved speaker diarization, In: ICASSP, pp. 4077–4080.

Gales, M.J.F., Kim, D.Y., Woodland, P.C., Chan, H.Y., Mrva, D., Sinha, R., Tranter, S.E., 2006. Progress in the CU-HTK broadcast news transcription system. IEEE Trans. Speech Audio Process. 14 (5), 1513–1525.

Galliano, S. et al., 2006. Corpus description of the ESTER evaluation campaign for the Rich Transcription of French Broadcast News, In: Proc. of Language Evaluation and Resources Conference.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., 1993. The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM. Linguist. Data Consort..

Garofolo, J. et al., 2002. NIST rich transcription 2002 evaluation: A Preview, LREC.

Gauvain, J.L., Lamel, L., Adda, G., 1998. Partitioning and transcription of broadcast news data, In: Proc. of ICSLP 4, Sydney, Australia, pp. 1335–1338.

Ghahramani, Z., Jordan, M.I., 1997. Factorial hidden Markov models. Mach. Learn. 29, 245–273.

Graff, D., 2001. TDT3 Mandarin Audio, Linguistic Data Consortium, Philadelphia.

Gravier, G., Betser, M., Ben, M., 2010. Audioseg: Audio Segmentation Toolkit, release 1.2, Product specification, Last updated 14 January 2010.

Griffiths, L., Jim, C., 1982. An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag..

Hain, T. et al., 1998. Segment generation and clustering in the HTK broadcast news transcription system, In: Proc. of DARPA Broadcast News Transcription and Understanding Workshop, pp. 133–137.

Han, K., Narayanan, S., 2008. Novel inter-cluster distance measure combining GLR and ICR for improved agglomerative hierarchical speaker clustering, In: Proc. of ICASSP, pp. 4373–4376.

Hansen, J.H.L. et al., 2005. SpeechFind: advances in spoken document retrieval for a national gallery of the spoken word. IEEE Trans. Speech Audio Process. 13 (5), 712–730.

Harb, H., Chen, L., 2006. Audio-based description and structuring of videos. Internat. J. Digital Libraries 6 (1), 70–81.

Heck, L., Sankar, A., 1997. Acoustic clustering and adaptation for robust speech recognition, In: Proc. of Eurospeech, Rhodes, Greece.

Hoshuyama, O., Sugiyama, A., Hirano, A., 1999. A robust adaptive beamformer for microphone arrays with a blocking matrix using coefficient-constrained adaptive filters. IEEE Trans. Signal Process..

Huang, C.H., Chien, J.T., Wang, H.M., 2004. A new eigenvoice approach to speaker adaptation, In: Proc. of Internat. Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, pp. 109–112.

Huang, Y., Vinyals, O., Friedland, G., Muller, C., Mirghafori N., Wooters, C., 2007. A fast-match approach for robust, faster than real-time speaker diarization", In: Proc. of IEEE Workshop on Speech Recognition and Understanding, pp. 693–698.

Huijbregts, M., Ordelman R., de Jong, F., 2007. Annotation of heterogeneous multimedia content using automatic speech recognition. In: Proc. of SAMT, vol. 4816, December 5–7 2007, Genova, Italy. Lecture Notes in Computer Science.

Hung, J., Wang, H., Lee, L., 2000. Automatic metric based speech segmentation for broadcast news via principal component analysis, In: Proc. of ICASLP, Beijing, China.

Hung, H., Huang, Y., Yeo, C., Gatica-Perez, D., 2008. Associating audio-visual activity cues in a dominance estimation framework, In: Proc. of CVPR Workshop on Human Communicative Behavior, Ankorage, Alaska.

The Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP) Research Institute, <http://www.idiap.ch/speech_processing.php>.

Imseng, D., Friedland, G., 2010. Tuning-robust initialization methods for speaker diarization. IEEE Trans. Audio Speech Language Process 18 (8), 2028–2037.

International Computer Science Institute Speech Research Group Berkeley. <http://www.icsi.berkeley.edu/groups/speech/>.

Istrate, D., Fredouille, C., Meignier, S., Besacier, L., Bonastre, J.-F., 2005. NIST RT'05 evaluation: preprocessing techniques and speaker diarization on multiple microphone meetings, In: Proc. of Machine Learning for Multimodal Interaction Workshop (MLMI), Edinburgh, U.K., pp. 428–439.

Izmirli, O., 2000. Using a spectral flatness based feature for audio segmentation and retrieval, (Abstract) In: Proc. of the Internat. Symposium on Music Information Retrieval (ISMIR2000), Plymouth, Massachusetts, USA.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ.

Janin, A. et al., 2003. The ICSI meeting corpus, In: Proc. of ICCASP, Hong Kong.

Jin, H., Kubala, F., Schwartz, R., 1997. Automatic speaker clustering, In: Proc. of DARPA Speech Recognition Workshop, pp. 108–111.

Jin, Q., Laskowski, K., Schultz, T., Waibel, A., 2004. Speaker segmentation and clustering in meetings, In: Proc. of the NIST Meeting Recognition Workshop, pp. 112–117.

Johnson, S., 1999. Who spoke when? automatic segmentation and clustering for determining speaker turns, In: Proc. of Eurospeech, Budapest, Hungary.

Johnson, D., Dudgeon, D., 1993. Array Signal Processing. Prentice Hall.

Johnson, S., Woodland, P., 1998. Speaker clustering using direct maximization of the MLLR adapted likelihood, In: Proc. of ICSLP 5, pp. 1775–1779.

Johnson, S.E., Woodland, P.C., 2000. A method for direct audio search with applications to indexing and retrieval, In: Proc. of ICASSP 3, Istanbul, Turkey, pp. 1427–1430.

Jothilakshmi, S., Ramalingam, V., Palanivel, S., 2009. Speaker diarization using autoassociative neural networks. Eng. Appl. Artificial Intell. 22 (4–5).

Juang, B., Rabiner, L., 1985. A probabilistic distance measure for hidden Markov models, AT& T, Technical Journal 64, AT&T.

Kaneda, Y., 1991. Directivity characteristics of adaptive microphone-array for noise reduction (amnor). J. Acoust. Soc. Jpn 12 (4), 179–187.

Kataoka, A., Ichirose, Y., 1990. A microphone array configuration for anmor (adaptive microphone-array system for noise reduction). J. Acoust. Soc. Jpn 11 (6), 317–325.

Kemp, T., Schmidt, M., Westphal, M., and Waibel, A., 2000. Strategies for automatic segmentation of audio data, In: Proc. of ICASSP 3, Istanbul, Turkey, pp. 1423–1426.

Kim, H., Elter, D., Sikora, T., 2005. Hybrid speaker-based segmentation system using model-level clustering, In: Proc. of ICASSP I, Philadelphia, USA, pp. 745–748.

Kingsbury, B., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech Commun. 25, 117–132.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors". Speech Comm. 52 (1), 12–40.

Kinnunen, T., Lee, K., Li, H., 2008. Dimension reduction of the modulation spectrogram for speaker verification, In: Proc. of Speaker and Language Recognition Workshop, IEEE Odyssey.

Koshinaka, T., Nagatomo K., Shinoda, K., 2009. Online speaker clustering using incremental learning of an ergodic hidden Markov model, In: Proc. of ICASSP, pp. 4093–4096.

Kotti, M., Benetos, E., Kotropoulos, C., 2006. Automatic speaker change detection with the Bayesian information criterion using MPEG-7 features and a fusion scheme, In: Proc. of the 2006 IEEE Internat. Symposium on Circuits and Systems, Greece.

Kotti, M., Martins, L.G.P.M., Benetos, E., Cardoso, J.S., Kotropoulos, C., 2006. Automatic speaker segmentation using multiple features and distance measures: a comparison of three approaches, In: Proc. of the 2006 IEEE International Conference on Multimedia and Expo, Toronto, Canada, pp. 1101–1104.

Kotti, M., Moschou, V., Kotropoulos, C., 2008. Speaker segmentation and clustering. Signal Process. 88 (5), 1091–1124.

Kotti, M., Benetos, E., Kotropoulos, C., 2008. Computationally efficient and robust BIC-based speaker segmentation. IEEE Trans. Audio Speech Language Process. 16, 920–933.

Kristjansson, T., Deligne, S., Olsen, P., 2005. Voicing features for robust speech detection, In: Proc. of ICSLP, Lisbon, Portugal.

Kubala, F. et al., 1997. The 1996 BBN byblos HUB-4 transcription system, In: Proc. of Speech Recognition Workshop, pp. 90–93.

Kuhn, R. et. al., 1998. Eigenvoices for speaker adaptation, In: Proc. of ICSLP, pp. 1771–1774.

Kuhn, R., Junqua, J.C., Nguyen, P., Niedzielski, N., 2000. Rapid speaker adaptation in eigenvoice space. IEEE Trans. Speech Audio Process. 8 (4), 695–707.

Kwon, S., Narayanan, S., 2002. Speaker change detection using a new weighted distance measure, In: Proc. of the Internat. Conf. on Spoken Language 4, USA, pp. 2537–2540.

Kwon, S., Narayanan, S., 2003a. A study of generic models for unsupervised online speaker indexing, In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop, pp. 423–428.

Kwon, S., Narayanan, S., 2003b. A method for online speaker indexing using generic reference models, In: Proc. of Eurospeech.

Kwon, S., Narayanan, S., 2004. Unsupervised speaker indexing using generic models. IEEE Trans. Speech Audio Process. 13 (5), 1004–1013.

Kwon, S., Narayanan, S., 2004a. Unsupervised speaker indexing using generic models. IEEE Trans. Speech Audio Process. 13, 1004–1013.

Kwon, S., Narayanan, S., 2004b. Speaker model quantization for unsupervised speaker indexing, In: Proc. of INTERSPEECH, pp. 1517–1520.

Lapidot, I., 2002. Self organizing maps with BIC for speaker clustering, IDIAP Research Report 02-60.

Lapidot, I., Guterman, H., 2001. Resolution limitation in speakers clustering and segmentation problems, In: Proc. of the 2001: A Speaker Odyssey, The Speaker Recognition Workshop, Chania, Greece, pp. 169–173.

Lathoud, G., McCowan, I.A., Odobez, J.M., 2004. Unsupervised location-based segmentation of multi-party speech, In: Proc. of ICASSP-NIST Meeting Recognition Workshop.

Leeuwen, D.A., Konecny, M., 2008. Progress in the AMIDA speaker diarization system for meeting data, In: Proc. of Multimodal Technologies for Perception of Humans: International Evaluation Workshops, Revised Selected Papers. vol. 4625 of LNCS, pp. 475–483.

The Laboratoire d'Informatique pour la Me' canique et les Sciences de l'Inge' nieur (LIMSI) Spoken Language Processing Group, <http://www.limsi.fr/TLP>.

Liu, D., Kubala, F., 1999. Fast speaker change detection for broadcast news transcription and indexing, In: Proc. of Eur. Conf. Speech Commun. Technol. III, Budapest, Hungary, pp. 1031–1034.

Liu, D., Kubala, F., 2003. Online speaker clustering, In: Proc. of ICASSP I, Hong Kong, China, pp. 572–575.

Liu, D., Kubala, F., 2004. Online speaker clustering, In: Proc. of ICASSP, pp. 333–336.

Lopez, J.F., Ellis, D.P.W., 2000. Using acoustic condition clustering to improve acoustic change detection on broadcast news, In: Proc. of ICSLP, Beijing, China.

Lu, L., Zhang, H.J., 2002. Real-time unsupervised speaker change detection," In: Proc. of ICPR, vol. 2, Quebec City, Canada.

Lu, L., Zhang, H., 2002. Speaker change detection and tracking in real-time news broadcast analysis, In: Proc. of the ACM Multimedia, France, pp. 602–610.

Lu, L., Zhang, H., 2005. Unsupervised speaker segmentation and tracking in real-time audio content analysis. Multimedia Syst. 10 (4), 332–343.

Lu, L., Zhang, H., Jiang, H., 2002. Content analysis for audio classification and segmentation. IEEE Trans. Speech Audio Process. 10 (7), 504–516.

Malegaonkar, A., Ariyaeeinia, A., Sivakumaran, P., Fortuna, J., 2006. Unsupervised speaker change detection using probabilistic pattern matching. IEEE Signal Process. Lett. 13 (8), 509–512.

Markov, K., Nakamura, S., 2007a. Never-ending learning system for online speaker diarization, In: Proc. of ASRU, pp. 699–704.

Markov, K., Nakamura, S., 2007b. Never-ending learning with dynamic hidden Markov network, In: Proc. of INTERSPEECH.

Markov, K., Nakamura, Satoshi, 2008. Improved novelty detection for online GMM based speaker diarization, In: Proc. of INTERSPEECH, Brisbane, Australia, pp. 363–366, September 22–26.

Marro, C. et al., 1998. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. IEEE Trans. Speech Audio Process..

Martin, A., Przybocki, M., 2001. Speaker recognition in a multi-speaker environment, In: Proc. of European Conference Speech Communications Technology, vol. 2, pp. 787–790.

McCowan, I., Marro, C., Mauuary, L., 2000. Robust speech recognition using near-field superdirective beamforming with post-filtering, In: Proc. of ICASSP 3, pp. 1723–1726.

McNeill, D., 2000. Language and Gesture. Cambridge University Press, New York.

Meignier, S., Merlin, T., 2010. Lium Spkdiarization: An Open Source Toolkit for Diarization". CMU SPUD Workshop.

Meignier, S., Bonastre, J., Igounet, S., 2001. E-HMM approach for learning and adapting sound models for speaker indexing, In: Proc. of A Speaker Odyssey, pp. 175–180.

Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.F., Besacier, L., 2005. Step-by-step and integrated approaches in broadcast news speaker diarization. Comput. Speech Lang. (20), 303–330.

Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.F., Besacier, L., 2006. Step-by-step and integrated approaches in broadcast news speaker diarization. Comput. Speech Language 20 (2-3), 303–330.

Meinedo, H., Neto, J., 2003. Audio segmentation, classification and clustering in a broadcast news task, In: Proc. of ICASSP, Hong-Kong, China.

Mesgarani, N., Shamma, S., Slaney, M., 2004. Speech discrimination based on multiscale spectro-temporal modulations, In: Proc. of ICASSP 1, Montreal, Canada, pp. 601–604.

Microsoft Audio Projects, <http://research.microsoft.com/users/llu/Audioprojects.aspx>.

Mirghafori, N., Wooters, C., 2006. Nuts and flakes: A study of data characteristics in speaker diarization, In: Proc. of ICASSP.

Mistral: Open Source platform for biometrics authentification, <http://mistral.univ-avignon.fr/en/index.html>.

Moattar, M.H., Homayounpour, M.M., 2009. Speaker clustering performance improvement using eigen-voice speaker adaptation, In: Proc. of 14th Internat. Computer Society of Iran Computer Conference (CSICC), pp. 501–506.

Moattar, M.H., Homayounpour, M.M., 2009. A Simple but efficient real-time voice activity detection algorithm, In: Proc. of 17th European Signal Processing Conf., (Eusipco), pp. 2549–2553.

Moh, Y., Nguyen, P., Junqua, J.-C., 2003. Toward domain independent clustering, In: Proc. of ICASSP II, pp. 85–88.

Moraru, D. et al., 2004. ELISA Nist RT03 broadcast news speaker diarization experiments, In: Proc. of Odyssey 2004, the Speaker and Language Recognition Workshop, Toledo, Spain.

Moraru, D., Besacier, L., 2003. Towards conversational model for speaker segmentation speech technology. Proc. Human-Comput. Dialogue Bucharest.

Moraru, D., Meignier, S., Besacier, L., Bonastre, J.-F., Magrin-Chagnolleau, Y., 2003. The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation, In: Proc. of ICASSP II, Hong Kong, pp. 89–92.

Moraru, D., Meignier, S., Fredouille, C., Besacier, L., Bonastre, J.-F., 2004. The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation, In: Proc. of ICASSP, Montreal, Canada.

Mori, K., Nakagawa, S., 2001. Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition, In: Proc. of ICASSP 1, pp. 413–416.

Muthusamy, Y.K. et al., 1992. The OGI multi-language telephone speech corpus, In: Proc. of ICSLP, pp. 895–898.

Neal, R.M., Hinton, G.E., 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. Learning in Graphical Models. The MIT Press, pp. 355–368.

Nguyen, P., Rigazio, L., Moh, Y., Junqua, J.C., 2002. Rich transcription 2002 site report. Panasonic speech technology laboratory (PSTL), In: Proc. Rich Transcription Workshop.

Nguyen, T.H., Li, H., Chng, E.S., 2009. Cluster criterion functions in spectral subspace and their application in speaker clustering, In: Proc. of ICASSP, pp. 4085–4088.

Nishida, M., Kawahara, T., 2003. Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion, In: Proc. of ICASSP 1, pp. 172–175.

NIST Fall Rich Transcription on meetings 2006 Evaluation Plan: 2006. <http://www.nist.gov/speech/tests/rt/rt2006/spring/docs/rt06s-meeting-eval-plan-V2.pdf>.

NIST Pilot Meeting Corpus website: 2006. <http://www.nist.gov/speech/testbeds/mr_proj/meeting_corpus_1>.

NIST Rich Transcription evaluations, 2006. In: Proc. of the 2nd International Conference on Image and Video Retrieval (CIVR'03), pp. 488–499.: <http://www.nist.gov/speech/tests/rt>.

H.J. Nock, G. Iyengar, C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," Lecture Notes in Computer Science, vol. 2728, 2003, pp. 565–570.

Noulas, A., Kros, B.J.A., 2007. On-line multi-modal speaker diarization. Proc. of ICMI. ACM, New York, NY, USA, pp. 350–357.

Noulas, A.K., Englebienne, G., Krose, B.J.A., 2009. Mutimodal speaker diarization. Proc. Comput. Vis. Image Understand..

Nwe, T.L., Sun, H., Li, H., Rahardja, S., 2010. Speaker diarization in meeting audio, In: Proc. of ICASSP, pp. 4073–4076.

Omar, M., Chaudhari, U., Ramaswamy, G., 2005. Blind change detection for audio segmentation" In: Proc. of ICASSP.

Otero, P.L., Fernandez, L.D., Mateo, C.G., 2010. Novel strategies for reducing the false alarm rate in a speaker segmentation system, In: Proc. of ICASSP, pp. 4970–4973.

Ouellet, P., Boulianne, G., Kenny, P., 2005. Flavors of Gaussian warping, In: Proc. of ICASSP, Lisbon, Portugal.

Pardo, J.M., Anguera, X., Wooters, C., 2006. Speaker diarization for multiple-microphone meetings using only between-channel differences. In: proc. of the Third Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2006). Washington DC, pp. 257–264.

Pardo, J.M., Anguera, X., Wooters, C., 2006. Speaker diarization for multiple distant microphone meetings: mixing acoustic features and inter-channel time differences, In: Proceedings of ICSLP, pp. 2194–2197.

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification, In: Proc. of ISCA Speaker Recognition Workshop, Odyssey, Crete, Greece.

Pellom, B.L., Hansen, J.H.L., 1998. Automatic segmentation of speech recorded in unknown noisy channel characteristics. Speech Comm. 25 (1–3), 97–116.

Pfau, T., Ellis, D.P.W., 2001. Hidden Markov model based speech activity detection for the ICSI meeting project, In: Proc. of Eurospeech.

Rao, R., Chen, T., 1996. Exploiting audio-visual correlation in coding of talking head sequences, In: Proc. of Internat. Picture Coding Symposium.

Reynolds, D.A., Torres-Carrasquillo, P., 2004. The MIT Lincoln Laboratories RT-04F diarization systems: Applications to broadcast audio and telephone conversations, In: Proc. of Fall 2004 Rich Transcription Workshop (RT04), Palisades, NY.

Reynolds, D.A., Torres-Carrasquillo, P., 2004. The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations, In: Proc. of Fall 2004 Rich Transcription Workshop (RT-04), Palisades, NY.

Rich Transcription Evaluation Project. National Institute of Technology (NIST), 2002–2009. <http://www.itl.nist.gov/iad/mig/tests/rt>.

The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan, <http://www.itl.nist.gov/iad/mig//tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf>.

Roch, M., Cheng, Y., 2004. Speaker segmentation using the map-adapted Bayesian information criterion, In: Proc. of Odissey, Toledo, Spain, pp. 349–354.

Rosca, J., Balan, R., Beaugeant, C., 2003. Multi-channel psychoacoustically motivated speech enhancement, In: Proc. of ICASSP.

Rougui, J., Rziza, M., Aboutajdine, D., Gelgon, M., Martinez, J., 2006. Fast incremental clustering of Gaussian mixture speaker models for scaling up retrieval in online broadcast, In: Proc. of ICASSP, Toulouse, France.

Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California, <http://sail.usc.edu/projectsIntro.php/>.

Sanchez-Bote, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J., 2003. A real-time auditory-based microphone array assessed with e-rasti evaluation proposal, In: Proc. of ICASSP.

Sankar, A., Beaufays, F., Digalakis, V., 1995. Training data clustering for improved speech recognition, In: Proc. of Eurospeech, Madrid, Spain.

Sankar, A., Weng, F., Stolcke, Z.R.A., Grande, R.R., 1998. Development of SRI's 1997 broadcast news transcription system, In: Proc. of DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, USA.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Shriberg, E., 2007. Higher-level features in speaker recognition speaker classification. Series Lecture Notes in Computer Science 4343, 241–259.

Shriberg, E., Stolcke, A., Baron, D., 2001. Observations on overlap: findings and implications for automatic processing of multi-party conversations. Proc. of Eurospeech, 1359–1362.

Sian Cheng, S., min Wang, H., 2003. A sequential metric-based audio segmentation method via the Bayesian information criterion, In: Proc. of Eurospeech, Geneva, Switzerland.

Siegler, M.A., Jain, U., Raj, B., Stern, R.M., 1997. Automatic segmentation, classification and clustering of broadcast news audio, In: Proc. of DARPA Speech Recognition Workshop, Chantilly, pp. 97–99.

Sinha, R., Tranter, S.E., Gales, M.J.F., Woodland, P.C., 2005. The Cambridge University March 2005 speaker diarisation system, In: Proc. of the European Conference on Speech Communication and Technology, pp. 2437–2440.

Siracusa, M., Fisher, J., 2007. Dynamic dependency tests for audio-visual speaker association, In: Proc. of ICASSP.

Sivakumaran, P., Fortuna, J., Ariyaeeinia, A., 2001. On the use of the Bayesian information criterion in multiple speaker detection, In: Proc. of Eurospeech, Scandinavia.

Solomonoff, A., Mielke, A., Schmidt, M., Gish, H., 1998. Clustering speakers by their voices, In: Proc. of ICASSP 2, Seattle, USA, pp. 757–760.

Stafylakis, T., Katsouros, V., Carayannis, G., 2009. Redefining the Bayesian information criterion for speaker diarisation, In: Proc. of INTERSPEECH.

Stafylakis, T. et. al., 2010. A new penalty term for the BIC with respect to speaker diarization, In: Proc. of ICASSP, pp. 4978–4981.

Stern, R., 1997. Specification of the 1996 Hub4 broadcast news evaluation, In: Proc. of DARPA Speech Recognition Workshop, (Chantilly, VA).

Stolcke, A., Friedland, G., Imseng, D., 2010. Leveraging speaker diarization for meeting recognition from distant microphones, In: Proc. of ICASSP, pp. 4390–4393.

Sturim, D., Reynolds, D., Singer, E., Campbell, J.P., 2001. Speaker indexing in large audio databases using anchor models, In: Proc. of ICASSP, Salt Lake City, USA.

Sun, H., Nwe, T.L., Ma B., Li, H., 2009. Speaker diarization for meeting room audio, In: Proc. of INTERSPEECH.

Sun, H. et al., 2010. Speaker diarization system for Rt07 and Rt09 meeting room audio, In: Proc. of ICASSP, pp. 4982–4985.

The Chair of Computer Science VI, Computer Science Department, Aachen University, <http://www-i6.informatik.rwth-aachen.de>.

Thyes, O., Kuhn, R., Nguyen, P., Junqua, J.C., 2000. Speaker identification and verification using eigenvoices, In: Proc. of ICSLP, Vol. 2, pp. 242–246.

Tranter, S., 2005. Two-way cluster voting to improve speaker diarization performance, In: Proc. of ICASSP, Montreal, Canada, 2005.

Tranter, S.E., Reynolds, D.A., 2006. An overview of automatic speaker diarization systems. IEEE Trans. Audio Speech Language Process. 14 (5), 1557–1565.

Tranter, S.E., Yu, K., Evermann, G., Woodland, P.C., 2004. Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech, In: Proc. of ICASSP, pp. 433–477.

Tritschler, A., Gopinath, R., 1999. Improved speaker segmentation and segment clustering using the Bayesian information criterion, In: Proc. of EuroSpeech, pp. 679–682.

Trueba-Hornero, B., 2008. Handling overlapped speech in speaker diarization, Master's thesis, Universitat Politecnica de Catalunya.

Tsai, W.H., Cheng, S.S., Wang, H.M., 2004. Speaker clustering of speech utterances using a voice characteristic reference space, In: Proc. of ICASLP, Jeju Island, Korea.

Vajaria, H., Islam, T., Sarkar, S., Sankar, R., Kasturi, R., 2006. Audio segmentation and speaker localization in meeting videos, In: Proc. of ICPR 2, pp. 1150–1153.

Valin, J., Rouat, J., Michaud, F., 2004. Microphone array post-filter for separation of simultaneous non-stationary sources, In: Proc. of ICASSP..

Vandecatseye, A., Martens J.P. et al., 2004. The cost278 pan-european broadcast news database, In: Proc. of LREC, Lisbon, Portugal.

van Leeuwan, D.A., 2005. The TNO speaker diarization system for NIST RT05s meeting data, In: Proc. of Machine Learning for Multimodal Interaction Workshop (MLMI), Edinburgh, UK, pp. 440–449.

Vescovi, M., Cettolo, M., Rizzi, R., 2003. A DP algorithm for speaker change detection, In: Proc. of the 8th European Conf. on Speech Communication and Technology, pp. 2997–3000.

Vijayasenan, D., Valente, F., Bourlard, H., 2007. Agglomerative information bottleneck for speaker diarization of meetings data, In: Proc. of ASRU, pp. 250–255.

Voitovetsky, I., Guterman, H., Cohen, A., 1997. Unsupervised speaker classification using self-organizing maps, In: Proc. of the IEEE

Workshop Neural Networks for Signal Processing, Amelia Island, USA, pp. 578–587.

Voitovetsky, I., Guterman, H., and Cohen, A., 1998. Validity criterion for unsupervised speaker recognition, In: Proc. of the First Workshop Text, Speech, and Dialogue, Brno, Czech Republic, pp. 321–326.

Wactlar, H., Hauptmann, A., Witbrock, M., 1996. News on-demand experiments in speech recognition, In: Proc. of ARPA STL Workshop.

Wang, D., Lu, L., Zhang, H.J., 2003. Speech segmentation without speech recognition, In: Proc. of ICASSP 1, Hong Kong, pp. 468–471.

Wang, W., Lv, P., Zhao, Q., Yan Y., 2007. A decision-tree-based online speaker clustering. In: Proc. of the 3rd Iberian conference on Pattern Recongnition and Image Analysis (IbPRIYA '07), Part I, Lecture Notes in Computer Science, vol. 4477, pp. 555-562.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. JASA 58, 236–245.

Wegman, S. et. al., 1997. Dragon system's 1997 broadcast news transcription system, In: Proc. of DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, USA.

Wölfel, M., Yang, Q., Jin, Q., Schultz, T., 2009. Speaker identification using warped MVDR cepstral features, In: Proc. of INTERSPEECH.

Woodland, P., Gales, M., Pye, D., Young, S., 1997. The development of the 1996 HTK broadcast news transcription system, In: Proc. of Speech Recorgnition Workshop, pp. 73–78.

Wooters, C., Huijbregts, M., 2008. The ICSI RT 2007 speaker diarization system, In: Proc. of the NIST Rich Transcription 2007 Spring Meeting Recognition Evaluation, RT07s LNCS vol. 4625.

Wooters, C., Fung, J., Peskin, B., Anguera, X., 2004. Toward robust speaker segmentation: the ICSI-SRI Fall 2004 diarization system, In: Proc. of Fall 2004 Rich Transcription Workshop (RT-04), Palisades, NY.

Wu, C.H., Hsieh, C.H., 2006. Multiple change-point audio segmentation and classification using an MDL-based Gaussian model. IEEE Trans. Audio Speech Language Process. 14 (2), 647–657.

Wu, T., Lu, L., Chen K., Zhang, H.J., 2003. UBM-based incremental speaker adaptation, In: Proc. of ICME 2, pp. 721–724.

Wu, T., Lu, L., Chen K., Zhang, H.J., 2003. UBM-based real-time speaker segmentation for broadcasting news, In: Proc. of ICASSP, pp. 193–196.

Wu, T., Lu, L., Chen, K., Zhang, H.J., 2003. Universal background models for real-time speaker change detection, In: Proc. of Internat. Conf. on Multimedia Modeling.

Yamaguchi, M., Yamashita, M., Matsunaga, S., 2005. Spectral cross-correlation features for audio indexing of broadcast news and meetings, In: Proc. of ICASLP.

Yoo, I.C., Yook, D., 2009. Robust voice activity detection using the spectral peaks of vowel sounds. ETRI J. 31 (4), 451–453.

Zamalloa, et al., 2010. Low latency online speaker tracking on the AMI corpus of meeting conversations, In: Proc. of ICASSP, pp. 4962–4965.

Zdansky, J., 2006. BINSEG: an efficient speaker-based segmentation technique, In: Proc. of INTERSPEECH, pp. 2186–2189.

Zelinski, R., 1988. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms, In: Proceedings of ICASSP 5, pp. 2578–2581.

Zhang, X., Hansen, J., Rehar, K., 2004. Speech enhancement based on a combined multichannel array with constrained iterative and auditory masked processing, In: Proc. of ICASSP.

Zhang, C., Yin, P., Rui, Y., Cutler, R., Viola, P., 2006. Boosting-based multimodal speaker detection for distributed meetings, In: Proc. of IEEE Internat. Workshop on Multimedia Signal Processing (MMSP).

Zhou, B., Hansen, J.H., 2000. Unsupervised audio stream segmentation and clustering via the Bayesian information criterion, In: Proc. of ICSLP 3, Beijing, China, pp. 714–717.

Zhou, B., Hansen, J.H.L., 2005. Efficient audio stream segmentation via the combined T2 statistic and the Bayesian information criterion. IEEE Trans. Speech Audio Process. 13 (4), 467–474.

Zhu, Y., Rong, X., 2003. Unified fusion rules for multisensory multi-hypothesis network decision systems. IEEE Trans. System Man Cybernet. 33 (4), 502–513.

Zhu, X., Barras, C., Meignier, S., Gauvain, J.L., 2005. Combining speaker identification and BIC for speaker diarization, In: Proc. of European Conf. on Speech Communications Technology, Lisbon, Portugal.

Zhu, X., Barras, C., Lamel, L., Gauvain, J.L., 2006. Speaker diarization: from broadcast news to lectures, In: Proc. of MLMI, pp. 396–406.

Zhu, X., Barras, C., Lamel, L., Gauvain, J.-L. 2008. Multi-stage speaker diarization for conference and lecture meetings, In: Proc. of Multimodal Technologies for Perception of Humans: Internat. Evaluation Workshops CLEAR 2007 and RT 2007, Revised Selected Papers. Berlin, Heidelberg: Springer-Verlag, pp. 533–542.

Zochova, P., Radova, V., 2005. Modified DISTBIC algorithm for speaker change detection, In: Proc. of ICSLP, Lisbon, Portugal, .