Robust Speaker segmentation and clustering for Meetings

(PhD Thesis Proposal)

Xavier Anguera Miró
TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
Campus Nord, mòdul D5
Jordi Girona, 1-3
08034 Barcelona, Spain
e-mail: xanguera@gps.tsc.upc.es

Thesis Advisor: Prof. Javier Hernando Pericás
TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
e-mail: javier@gps.tsc.upc.es

March 2005

Contents

1 Introduction

In this work a speaker clustering system for the meetings environment is proposed for a thesis topic.

In order to define speaker clustering we need to put it in the context of speaker segmentation and acoustic change detection, as they are sometimes confused and used alternatively.

Given an audio source, we will be performing an acoustic change detection if our aim is to simply locate the points in time where there is a change in the kind of music played, the background noise or the speaker that is talking. We can see a SAD (Speech Activity Detector) as a simple acoustic change detector.

When we focus on detecting changes of speakers talking, isolating them from the background conditions, we are performing a speaker segmentation. It is normally aimed to tasks where the discrimination of speakers is the primary concern.

Finally, speaker clustering systems try to group together all speech segments that belong to one particular speaker. This can be done within a single recording or in a whole database (different background and recording conditions make this task particularly difficult). Each speaker is not assigned a true identity, only a unique identifier. It is a duty of speaker identification systems to link each identifier to each speaker identity.

Speaker segmentation and clustering is nowadays one of the hot topics in transcription and indexing of speech signals, although it has been used as an important input for speech recognition and speaker identification for a few years.

On one hand, the importance of speaker segmentation in speech recognition relies on the fact that the decoder part in recognition systems works best with short audio segments containing always complete words. Speaker segmentation in this case acts as a butcher, cutting the signal in convenient places. On the other hand, for speaker identification and for acoustic modelling adaptation in speech recognition, it is important to be able to train or adapt each acoustic model from speaker homogeneous data. This data is taken from the output of speaker clustering systems.

Provably the latest application, but yet the most challenging, for a speaker segmentation and clustering system is audio indexing and automatic transcription. In these systems we are trying to accurately answer the question "Who Spoke When?" in order to obtain an accurate time sequence of all the speaker turns in a recording. This is the most challenging task as the output of the segmentation system goes directly into the output transcript and because small errors in this segmentation can affect much more the credibility and usability of the system.

In this third application, an initial strong emphasis was put into Broadcast News segmentation systems([7]), in order to be able to Index the data

2

created by radio and TV stations in an automatic matter.Nowadays all these transmissions are stored to make them available for future reference. Such recordings will undergo several levels of manual labelling to improve their future usability, but this is normally insufficient and lacks of consistency.

Speaker clustering allows to automatically and consistently segment the recordings. Together with speaker identification, speech recognition and other speech technologies a complete indexing can be produced.

In the later years the use of speech technologies has gained a lot of interest in the Meetings environment ([1]; [2]; [3]; [4]), where the goal is double: help to the good development of a meeting (during the meeting) with the use of speech and video capabilities, and produce a transcription/indexation of the meeting that allows its offline browsing or automatic creation of meeting minutes.

In general, a meeting is defined to be a group of people that interact with each other in order to exchange information about a common interest. These people do not need to be in the same physical place, but regarding speaker segmentation and clustering we consider that they are.

Two main meetings environments are being studied within the projects described above: the business meeting (focus of the AMI project) describes a big round table with people sitting around it and all interacting towards the definition/design of a particular object. On the other hand, the lecture meeting (focus of the CHIL project) defines a lecturer giving a presentation in front of a small audience with a questions and answers turn at the end, where the attendants get to intervene.

The main goal of this thesis is to build a speaker segmentation and clustering system for a meetings environment starting from current broadcast news clustering technology. The system will focus on the particularities of each system to make use of the advantages and reduce the effect of the issues present in the transition from one technology to the other. In doing so I intend to bring new ideas into speaker clustering, beneficial both for Meetings and Broadcast News environments

The resulting system is aimed to be as robust as possible to any room distribution, microphones amount and placement and kind of meeting.

The remaining sections of this proposal go as follows. Section 2 presents the state of the art of speaker segmentation and speaker clustering and chapter 3 presents a work plan for the achievement of the proposed goals.

2   State of the Art

In this chapter it is presented the most significant work in speaker segmentation and speaker clustering done up to nowadays, putting emphasis in the application of such techniques in the meetings environment.

The speaker segmentation techniques are an area of work on their own,

but are many times also used as a first step towards speaker clustering, although some other methods have been proposed that don't use an initial segmentation. In the following two sections we present the main approaches to speaker segmentation and to speaker clustering, showing their similarities.

2.1   Speaker segmentation

Speaker segmentation has sometimes been referred as speaker change detection and is closely related to acoustic change detection. For a given speech/audio segment it strives to find the acoustic change points due to change of speaker and/or background conditions.

It belongs to the pattern classification family, where we try to find categorical (discrete) labels/classes for continuous observations of speech and by doing so we find the boundaries between them. Speech recognition and speaker clustering are also pattern classification problems, and as such, they all need to work on a feature set that represents well the acoustic data and define a distance measure/method to assign each feature vector to each class.

Speaker segmentation has traditionally been performed using MFCC features with very few differences to the speaker verification or speech recognition tasks. The method used to find the change point is what distinguishes the different systems. In the bibliography ([5]; [6]; [7]; [8]; [9]) these are usually separated into three groups:

- Silence-based: The input stream is segmented in the silence parts. Two different techniques have been used: energy based systems use an energy detector and set the change points to the minimums than exceed a threshold ([6]; [10]; [11]). In [12] the MAD (Mean absolute deviation statistic), which measures the variability in energy within segments, is used instead.

  On the other hand, decoder-guided segmenters run a recognition system and obtain the change points from the silence locations detected ([13]; [14]; [15]; [16]) normally constraining their minimum duration. Some of these systems use some extra information from the decoder, such as gender labels ([17]). The output has normally been used as an input to recognition systems, but not for indexing or Diarization as there is not a clear relationship between the existence of a silence in a recording and change of speaker.

- Model based segmentation: Initial models (for example GMM's) are created for a closed set of acoustic classes (telephone-wideband, male-female, music-speech-silence and combinations of them) by using training data. The audio stream is classified by ML (Maximum Likelihood) selection using these models ([18]; [6]; [19]; [20]). The boundaries between models become the segmentation change points. Using trained

models has a robustness problem as they do not generalize for unseen data. In later years some clustering systems make use of an ML decoding of evolutive models that look for the optimum acoustic change points. In [21] and [22] it is done bottom-up and in [23] and [24] is done top-bottom. In [25] a prior segmentation in pretrained models is combined with an evolutive segmentation. Finally, in [26] SVMs (Support Vector Machines) are used as a classifier instead of ML for pretrained models.

- Metric based segmentation: The audio stream is segmented evaluating a metric between two neighboring audio segments (of the same of different length). The change points are the local maxima of the metric of the segments around it. There are many possible metrics to use. based on popularity I propose to separate them in two groups:

  - BIC related: The Bayesian Information Criterion is provably the most extensively used segmentation and clustering metric due to simplicity and effectivity. It is a likelihood criterion penalized by the model complexity (amount of parameters to train).

    Given $\mathcal{X}_1$ and $\mathcal{X}_2$ two adjacent audio segments, modelled with multi-gaussian processes $\mathcal{N}_1(\mu_1, \sigma_1)$ and $\mathcal{N}_2(\mu_2, \sigma_2)$ respectively, and $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$, modelled by $\mathcal{N}(\mu, \sigma)$.

    We define the BIC value for each one of the segments as:

    $$BIC(\mathcal{N}_i) = log\mathcal{L}(\mathcal{X}_i, \mathcal{N}_i) - \lambda\frac{1}{2}\#(\mathcal{N}_i)log(N) \qquad (1)$$

    Being $log\mathcal{L}(\mathcal{X}_i, \mathcal{N}_i)$ the likelihood of the data given the considered model, being λ a free design parameter dependent on the data being modelled and N the amount of frames in the considered segment.

    When evaluating wether a change point occurs between both segments we evaluate the hypothesis that $\mathcal{X}$ better models the data versus the hypothesis that $\mathcal{X}_1 + \mathcal{X}_2$ does instead, by computing:

    $$\Delta BIC(i) = -R(i) + \lambda P \qquad (2)$$

    Where P is the penalty term, which for a full covariance matrix is $P = \frac{1}{2}(p + \frac{1}{2}p(p+1))log(N_\mathcal{X})$, p being the dimension of the acoustic space. We can write the factor R(i) as:

    $$R(i) = \frac{N_\mathcal{X}}{2}log|\Sigma_\mathcal{X}| - \frac{N_{\mathcal{X}_1}}{2}log|\Sigma_{\mathcal{X}_1}| - \frac{N_{\mathcal{X}_2}}{2}log|\Sigma_{\mathcal{X}_2}| \qquad (3)$$

    The criterion was first introduced by G. Schwarz ([27];[28]) and later started to be used for acoustic change detection by Chen and

5

Gopalakrishnan ([8];[29];[7]). In the initial formulation a tunable parameter λ is introduced to adjust the penalty term which is a hidden threshold to the BIC difference. Several works address the problem of selecting or tuning this parameter ([30]; [31]; [32]; [33]; [34]; [35]). In [36] the penalty term is cancelled by carefully selecting the amount of features in each model.

BIC presents in general a problem when there is a big mismatch between the length of the adjacent windows. Some works have successfully applied normalizations to the original formula, either to the penalty term ([9]) or to the overall value ([37]) to reduce its effect.

Several implementations of BIC have been proposed in the literature. Initially [8] proposed a multiple changing points detection algorithm, and later [30], [38], [39], [40], [41], [42]. They all propose the use of a growing window with inner variable length analysis segments system to iteratively find the changing points. In [30] proposes some heuristics to make the algorithm faster and to focus on detecting very short changes. In [38], [41] and [42] speedups are proposed in computing the mean and variances of the models. In [43] a MAP-adapted version of the models is presented, which allows for shorter change points to be found

Even with the efforts to speed up the processing of BIC, it is very costly computationally to use it analyzing all the signal with a decent resolution. Some two-pass methods appeared which first do a quick selection of possible change points using another acoustic metric and then use BIC to assert or reject these changes. An important step in this direction is taken with DISTBIC ([31], [32], [44]) where the GLR (Generalized Likelihood Ratio)(see (refGLR)) is used prior to BIC. Also in this direction are [45], proposing the to use Hotelling's T^2 statistic, [40] using K-L2 (Kullback-Leibler) distance (see (5)) and [35] using a measure called NLLR (Normalized Log Likelihood Ratio).

– Other distances: Many other distances have been proposed in order to locate the change points in an audio stream, either combined in a multi-step segmentation (as in the previous cases) or stand alone.

  In [46] the GLR is used to segment the signal into speaker turns towards doing speaker verification. Given $\mathcal{X}_1$ and $\mathcal{X}_2$ as defined above, the GLR is defined in (4).

$$GLR = \frac{\mathcal{L}(\mathcal{X}, \mathcal{N}(\mu, \sigma)}{\mathcal{L}(\mathcal{X}_1, \mathcal{N}_1(\mu_1, \sigma_1))\mathcal{L}(\mathcal{X}_2, \mathcal{N}_2(\mu_2, \sigma_2))} \qquad (4)$$

  Which comes to be a ΔBIC without the penalty term. In [47]

and [48] a variation of the GLR distance is presented for speaker change detection, being called the Gish-distance from then on.

Another well used distance measure is the K-L and K-L2 distances ([49]; [50]): given two random variables A, B, the K-L distance is defined as

$$KL(A;B) = E_A(log\frac{P_A}{P_B})$$

Where E_A is the expected value with respect to the PDF of A. In order to symmetrize the K-L distance, the K-L2 is defines as:

$$KL2(A;B) = KL(A;B) + KL(B;A)$$

If we consider the acoustic segments $\mathcal{X}_1$ and $\mathcal{X}_2$ to be observations of the random variables A and B, we can model them with with multi-gaussian PDFs, and we can write:

$$KL2(A;B) = \frac{\sigma_A^2}{\sigma_B^2} + \frac{\sigma_B^2}{\sigma_A^2} + (\mu_A - \mu_B)^2(\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}) \qquad (5)$$

In [50] the acoustic vectors dimensionality is reduced with a PCA (Principal Component Analysis) and K-L2 or other distances are used to find the segmentation change points. In [6] a comparison is made between the K-L2, GLR and other methods.

In [51] a metric called XBIC is presented that compares how similar are both segments. Continuing with the same notation as before:

$$XBIC(\mathcal{X}_1;\mathcal{X}_2) = P(\mathcal{X}_1,\mathcal{N}_2(\mu_2,\sigma_2)) + P(\mathcal{X}_2,\mathcal{N}_1(\mu_1,\sigma_1))$$

In [52], [53] and [54] they obtain the segment's PDF's by MAP (Maximum a Posteriori) adaptation from a UBM and use a measure called divergence shape distance, defined in a similar fashion than [47]. In [55] they also use the divergence shape distance where the speaker models are incrementally updated, similar to [26] and [56] where extra measures are proposed.

In [57] the features in the adjacent segments are clustered into 3 subsets and a distance is computed between each pair, the biggest one is considered in evaluating possible change points.

In some works the adjacent scrolling windows architecture is substituted by other systems. In [58] GLR is used to segment 2 speakers conversations where assumes that the conversation is initiated by one of the speakers. In [33] uses a frame by frame VQ (Vector Quantization) distortion measure ([59]) and compares results to GLR and BIC techniques. In [60] a two pass algorithm is proposed in the same fashion as [31]. In this case a decoder-guided change points output is refined by a penalized GLR (BIC alike).

7

2.2 Speaker Clustering

In speaker clustering the aim is to group together with the same identifier all frames belonging to the same speaker in an audio stream. Although in some cases there exists some information about the number of speakers or their identity, we will only consider systems that propose solutions to the blind speaker clustering problem where there is no information at all.

Clustering results used for acoustic models adaptation tend to group together acoustically homogeneous speakers (even if they are different speakers) in order to have more data for adaptation. Systems intended for indexing (also called diarization systems) strive to obtain an accurate distinction between speaker.

Most of the state of the art systems use hierarchical schemes for this purpose. Speech segments are split or merged until the optimum number of speakers is reached In doing so, all systems need to define:

1. A distance between clusters/segments to determine acoustic similarity between them. A distance matrix is created with the distance from any possible pair.

2. A stopping criteria to stop the mergin/splitting at the optimum amount of clusters. In the case of real-time systems the distance measure threshold defines the stopping criteria as the number of final clusters depend directly on it.

Depending on the starting point we can distinguish:

- Bottom-up clustering: This is by far the most used approach as systems typically perform a clustering over the segmentation output. Normally the matrix distance between all segments is computed and the closest pair is merged iteratively until the stopping criteria is met.

  The most commonly used distance and stopping criteria is again BIC. BIC was initially proposed for clustering in [8] and [29]. At each iteration, the pair with bigger ΔBIC value is merged. The process finishes when all pairs have a ΔBIC< 0. Some later works ([7]; [30]; [17]; [41]; [61]) propose modifications to the penalty term and differences in the segmentation setup.

  One of the earliest works in speaker clustering was proposed in [62], using the Gish distance (see [47]) for distance matrix and as a stop criterion, the minimization of a penalized version (to avoid over-merging) of the within-cluster dispersion matrix.

  At the same time, in [49] the K-L2 divergence distance is used as a distance metric and stopping criterion by setting a merging threshold. Its results are seen to work better than the Mahalanobis distance.

In [63] uses GLR and K-L2 as distance matrices and iteratively merges clusters until it maximizes the cluster purity. The same stopping criterion is used in [64], where several methods are presented to create a different reference space for the acoustic vectors that better represents similarities between speakers. The cosine measure is used as a distance matrix.

In [65] and [66] a symmetric relative entropy distance is used and an empiric threshold is defined as stopping criterion. The distance is computed at a segment level (all segments from all clusters) and only the biggest value is shown to work the best.

Some works integrate the segmentation with the clustering by using a model-based segmentation. This is the case in [21], [22] and [67] where an initial segmentation is used to train speaker models that iteratively decode and retrain on the acoustic data. A threshold-free BIC metric ([36]) is used to merge the closest clusters at each iteration and as stopping criterion.

In [68] a two passes clustering is proposed. An initial clustering is obtained using a GLR distance matrix over equal length segments with agglomerative clustering until a known amount of speakers is reached. The second pass involves training of speaker models and iterative decoding-training on the data until the total likelihood converges.

In [69] a novel approach to speaker clustering is proposed using speakers triangulation to cluster the speakers. The probability of one segment given a model created from another segment is used as a distance metric.

- Top-down clustering: Fewer are the systems that start from one cluster and iteratively split until the stopping criterion is met. Two main works in this area have been found. In [70], [71] and [17] they propose a top-down system where previously created segments are the units used for clustering. They use the Arithmetic Harmonic Sphericity (AHS) (see [72]) and the Gaussian Divergence as splitting metric, and iterate to maximize the Maximum Likelihood Linear Regression (MLLR) Adaptation Likelihood. After each merge they use a measure similar to cross entropy to allow close clusters to merge.

In [23] and [24] an initial cluster is trained with all the acoustic data available. Iterative decoding-MAP adaptation is performed where new clusters are split using the likelihood over a 3 seconds window as a splitting measure. The overall likelihood is used as a stopping criterion. In [24] a repository model is further introduced to improve the purity of the created segments.

2.3 Segmentation and Clustering in Meetings

Within the NIST 2004 Spring Rich Transcription Evaluation ([73]) speaker diarization was evaluated in meetings recording in two different conditions: Multiple Distant Microphones (MDM) and Single Distant Microphones (SDM). This is the first time that this task was performed for meetings environment. In the 2002 NIST evaluation only the SDM task was done.

Following are the approaches that the participants proposed for the MDM case:

- Macquarie University in [74] proposes to perform clustering only on the most centrally located microphone, information which was available in the evaluation.

- The ELISA group in [75] proposes a two-axis merging strategy. An horizontal merging consists on collapsing and resegmentation of the clustering output of their two expert systems, based on BIC and EHMM, as was done in the RT'03 evaluation ([76]). This is done for each individual MDM channel. The vertical merging unifies all the individual channels into one result with an iterative process that searches for the longest speaker intervention.

- Carnegie Mellon University (CMU) in [77] presents a clustering scheme based on GLR distance and BIC stopping criterion. In order to obtain the initial segmentation of the data, first a Speech Activity Detection (SAD) is done over all the channels, then the resulting segmentations are collapsed into one and the best channel is selected for each one, based on energy and SNR. Speaker change detection is done using a scrolling GLR measure over large segments.

  In the point of view of speaker segmentation and clustering both tasks (meetings and broadcast news) present many differences to be taken into account and explored in order to adapt the technology used in Broadcast News to Meetings environments. We can see some of them in table 1.

3 Databases

In order to research and test speaker diarization systems for Meetings there need to be meetings databases accurately transcribed into speaker segments. Nowadays there exist a few databases generally available and a few more are being recorded and transcribed as I write this.

- ICSI Meetings Corpus ([78]; [79]): 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 mounted on a mock PDA.

Meetings Environment	Broadcast News Environment
Reduced amount of speakers, limited by the capacity of the room	totally unknown amount of speakers
There are neither music or commercials	there can be commercials and background music with speech
There are impulsive noises (doors shut down, pens fall, speakers touch their mics...)	different background conditions occur when reporting from the field
All recordings take place in the same setting	recordings alternate between studio and field (different bandwidth conditions).
Major use of spontaneous speech, with more silences and filling words/sounds	Much more scripted speech with professional narrators.
The average speaker turn can be very small (for example yes/no answers)	The average speaker turn is longer
Normal existence of overlapping regions where two or more people speak at the same time	normally there is no overlapping speech
The recordings are performed using several microphones	only one channel is available
The far-field channels (microphones in the meeting table) regularly have worse quality than closer mics	The speech quality is the regular broadcasting quality.

Table 1: *Main differences between Meetings and Broadcast News recordings*

- CMU Meeting Corpus ([80]; [81]) : 104 meetings of an average duration of 60 minutes with 6.4 participants (in average) per meeting (only 18 meetings are publicly available through LDC). They are focused on a given scenario or topic, changing from meeting to meeting. Initial meetings had 1 omnidirectional microphone, newer ones have 3 omnidirectional tabletop microphones.

- NIST Pilot Meeting Corpus([82]): Consists on 19 meetings with a total of about 15 hours. Several meeting types are proposed to the attendants. Recordings for MDM are done using 3 omnidirectional table-top microphones and one 4x circular directional microphone.

- Meetings in the NIST Meetings evaluations: In 2002 NIST (cite [83]) started an evaluation topic centered on the transcription of speech in meetings. In that year NIST made public fully hand-transcribed meetings from CMU (Carnegie Mellon University), ICSI and LDC (Linguis-

tic Data Consortium). In 2002 diarization for MDM was not evaluated. Again in 2004 (see [73]) a second evaluation was performed, this time including diarization. This year meetings were transcribed from ICSI, LDC, CMU and NIST.

- CHIL Corpus: Recordings are projected to be done in several meeting rooms in different partners locations. Some tabletop microphones record the MDM data.

- AMI corpus: It is projected to record about 100 hours of 4 participants meetings. These are split into two main groups: real meetings and scenario-based meetings (where people are briefed to talk about a particular topic). One circular array of 8 microphones is centrally located in the table.

4   Objectives and Work Plan

As it was explained in the introduction, the main goal of this thesis is to build a full speech clustering system for the meetings environment which takes advantage and improves the existing technology in speaker segmentation and clustering while exploring the advantages present in meetings data.

Some research and development has already been done in this direction. On one hand, a system to combine multiple input channels into an enhanced one has already been implemented using speech beamforming techniques, although some further research will be done in this area.

On the other hand some experimentation has been done with several speaker clustering methods in order to find a suitable one as baseline technology for meetings. The system presented in [22] has been selected because of its robustness to background changes and for being train-free when switching to new environments.

This system uses a model-based segmentation mixed with a BIC-like based clustering (both for a distance matrix and stopping criterion). It was designed for diarization of Broadcast News systems and has been used in recent evaluations ([67]).

Initially a fixed number of clusters is created by assigning equal length segments to each of them. Gaussian Mixture Models (GMMs) are trained and various iterations of Viterbi decoding and models retraining take place to obtain acoustically homogeneous models. Then a BIC-alike metric (presented in [36]) is computed between each cluster pair and the pair with biggest value is merged, followed by another set od decoding-retraining. This is performed iteratively until no positive values are found.

In figure 1 we can see a general block diagram with the proposed system implementation. These proposed blocks are explained in more detail below.
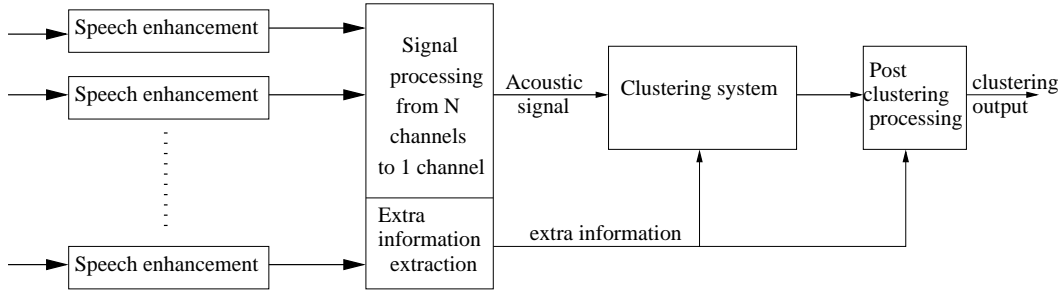
Figure 1: *Meetings diarization system blocks diagram*

- Improvements to the Broadcast News (BN) diarization system.

  By using a system that is robust to the change from Broadcast News to Meetings environments we ensure that improvements that benefit one kind will provably also benefit the other. I intend to improve the basic Broadcast news system in several ways:

  - Meetings/shows flakiness (1-6 months)

    One such task is to reduce the score variability between shows/meetings. It has been seen in the previous evaluations for meetings and BN ([73], [84]) that similar data produces very different Diarization Error Rates (DER) and that slight modifications on the system's parameters can also affect greatly in these scores. I believe that clusters purity plays an important role in our system's results. Some techniques will be applied to ensure that all segments in a cluster belong to the same speaker. In Broadcast News a step towards this direction is the detection of commercials in the shows. These are present in all shows and can affect the quality of the systems if not taken into consideration.

  - Computational reduction (3-6 months)

    Another common task is the computational cost reduction of the system, aiming towards execution in real time with no loss on the DER. To achieve it, methods to reliably merge more than one cluster at a time have to be explored, and the distance matrix computation has to be sped up.

- Adaptation of the diarization system to meetings

  In order to adapt the Broadcast News system to the peculiarities of meetings we have to take into account the differences between the two environments, as shown in table 1. At the diarization system level I intend to study:

  - Turn duration reduction (1-3 months)

13

the shorter duration of speaker turns in meetings causes a problem for the current system as it is optimized to work with BN durations. The GMM models need to be changed to allow shorter segments while avoiding models undertraining. This will be done by studying possible changes in the GMM topology, and how to enforce speaker changes when outside information is available.

– Multichannel information treatment (3-6 months)

Any information that can be provided to the system by the N-channels beamformer will need to be incorporated to help the diarization. Such information includes, but is not limited to, speaker overlap detection (where more than one speaker talks at the same time, which causes many errors in BN systems limited to one speaker turns).

- Multi-microphone processing for signal enhancement and information retrieval (6 months).

The system that has been implemented applies a delay&sum to all MDM channels, estimating the delay of arrival (DOA) of each channel with respect ro a reference channel. Then this delay is applied to the channels and they are summed to obtained the output signal. This is done over a small scrolling window in order to be sensitive to speaker/source changes. Future work will involve finding techniques to improve the estimation of such delay, making it robust to noise and silence parts. Also important is the extraction of extra information from the channels, such as the most important channel at each time, whether exists overlapping speech at a certain time, silence activity detection (SAD) and others.

As seen in figure 1, such information is expected to be used either in the diarization system or in a postprocessing module, after clustering.

- Evaluation of the system by participating in Meetings evaluations (1 month).

In order to test the proposed system I intend to participate in the Meetings evaluations campaigns proposed by NIST every Spring. The baseline system and a few proposed improvements will be presented to the evaluation in May 2005 (see [73]) and I hope to be able to test an improved systems in the 2006 evaluation.

I also intend to help the Broadcast News diarization team in their participation on the fall 2005 NIST evaluation (provided there is one).

References

[1] Augmented multiparty interaction (ami) website. [Online]. Available: http://www.amiproject.org

[2] Computers in the human interaction loop (chil) website. [Online]. Available: http://chil.server.de

[3] Interactive multimodal information management (im2) website. [Online]. Available: http://www.im2.ch

[4] Multimodal meeting manager (m4) website. [Online]. Available: http://www.m4project.org

[5] J. Ajmera, "Robust audio segmentation," Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne, 2004.

[6] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," in *ICASSP'00*, Istanbul, Turkey, 2000, pp. 1423–1426.

[7] S. S. Chen, M. J. F. Gales, R. A. Gopinath, D. Kanvesky, and P. Olsen, "Automatic transcription of broadcast news," *Speech Communication*, vol. 37, pp. 69–87, 2002.

[8] S. Shaobing Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, Feb. 1998.

[9] L. Perez-Freire and C. Garcia-Mateo, "A multimedia approach for audio segmentation in tv broadcast news," in *ICASSP'04*, Montreal, Canada, May 2004, pp. 369–372.

[10] S. Wegmann, F. Scattone, I. Carp, L. Gillick, R. Roth, and J. Yamron, "Dragon system's 1997 broadcast news transcription system," in *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, USA, 1998.

[11] H. Wactlar, A. Hauptmann, and M. Witbrock, "News on-demand experiments in speech recognition," in *ARPA STL Workshop*, 1996.

[12] M.-H. Siu, G. Yu, and H. Gish, "An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers," in *ICASSP-92*, vol. 2, San Francisco, USA, 1992, pp. 189–192.

15

[13] F. Kubala, H. Jin, S. Matsoukas, L. Gnuyen, R. Schwartz, and J. Machoul, "The 1996 bbn byblos hub-4 transcription system," in *Speech Recognition Workshop*, 1997, pp. 90–93.

[14] P. Woodland, M. Gales, D. Pye, and S. Young, "The development of the 1996 htk broadcast news transcription system," in *Speech Recorgnition Workshop*, 1997, pp. 73–78.

[15] T. Hain, S. Johnson, A. Turek, P. Woodland, and S. J. Young, "Segment generation and clustering in the htk broadcast news transcription system," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 133–137.

[16] J. F. Lopez and D. P. W. Ellis, "Using acoustic condition clustering to improve acoustic change detection on broadcast news," in *ICSLP-00*, Beijing, China, 2000.

[17] S. Tranter and D. Reynolds, "Speaker diarization for broadcast news," in *ODISSEY'04*, Toledo, Spain, May 2004.

[18] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *ICSLP-98*, vol. 4, Sidney, Australia, 1998, pp. 1335–1338.

[19] R. Bakis, S. Chen, P. Gopalakrishnan, and R. Gopinath, "Transcription of broadcast news shows with the ibm large vocabulary speech recognition system," in *Speech Recognition Workshop*, 1997, pp. 67–72.

[20] A. Sankar, F. Weng, Z. R. A. Stolcke, and R. R. Grande, "Development of sri's 1997 broadcast news transcription system," in *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, USA, 1998.

[21] J. Ajmera, H. Bourlard, and I. Lapidot, "Improved unknown-multiple speaker clustering using hmm," IDIAP, Tech. Rep., 2002.

[22] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *ASRU'03*, US Virgin Islands, USA, Dec. 2003.

[23] S. Meignier, J.-F. Bonastre, and S. Igournet, "E-hmm approach for learning and adapting sound models for speaker indexing," in *A speaker Oddissey*, Chania, Crete, 2001, pp. 175–180.

[24] X. Anguera and J. Hernando, "Evolutive speaker segmentation using a repository system," in *ICSLP'04*, Jeju Island, Korea, Oct. 2004.

[25] S. Meignier, D. Moraru, C. Fredouille, L. Besacier, and J.-F. Bonastre, "Benefits of prior acoustic segmentation for automatic speaker segmentation," in *ICASSP-04*, Montreal, Canada, 2004.

[26] L. Lu, S. Z. Li, and H.-J. Zhang, "Content-based audio segmentation using support vector machines," in *ACM Multimedia Conference*, 2001, pp. 203–211.

[27] G. Schwarz, "A sequential student test," *The Annals of Statistics*, vol. 42, no. 3, pp. 1003–1009, 1971.

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

[29] S. S. Chen and P. Gopalakrishnan, "Clustering via the bayesian information criterion with applications in speech recognition," in *ICASSP'98*, vol. 2, Seattle, USA, 1998, pp. 645–648.

[30] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Eurospeech'99*, 1999, pp. 679–682.

[31] P. Delacourt and C. J. Wellekens, "Distbic: A speaker-based segmentation for audio data indexing," *Speech Communication: Special Issue in Accessing Information in Spoken Audio*, vol. 32, pp. 111–126, 2000.

[32] P. Delacourt, D. Kryze, and C. J. Wellekens, "Detection of speaker changes in an audio document," in *Eurospeech-1999*, Budapest, Hungary, 1999.

[33] K. Mori and S. Nakagawa, "Speaker change detection and speaker clustering using vq distorsion for broadcast news speech recognition," in *ICASSP-01*, vol. 1, Salt Lake City, USA, May 2001, pp. 413–416.

[34] J. F. Lopez and D. P. W. Ellis, "Using acoustic condition clustering to improve acoustic change detection on broadcast news," in *ICSLP-00*, Beijing, China, 2000.

[35] A. Vandecatseye, J.-P. Martens, *et al.*, "The cost278 pan-european broadcast news database," in *LREC'04*, Lisbon, Portugal, May 2004.

[36] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," IDIAP, Tech. Rep., 2003.

[37] A. Vandecatseye and J.-P. Martens, "A fast, accurate and stream-based speaker segmentation and clustering algorithm," in *Eurospeech'03*, Geneva, Switzerland, 2003, pp. 941–944.

[38] P. Sivakumaran, J. Fortuna, and A. Ariyaeeinia, "On the use of the bayesian information criterion in multiple speaker detection," in *Eurospeech'01*, Scandinavia, Sept. 2001.

[39] S. sian Cheng and H. min Wang, "A sequential metric-based audio segmentation method via the bayesian information criterion," in *Eurospeech'03*, Geneva, Switzerland, 2003.

[40] L. Lu and H.-J. Zhang, "Real-time unsupervised speaker change detection," in *ICPR'02*, vol. 2, Quevec City, Canada, 2002.

[41] M. Cettolo and M. Vescovi, "Efficient audio segmentation algorithms based on the bic," in *ICASSP'03*, 2003.

[42] M. Vescovi, M. Cettolo, and R. Rizzi, "A dp algoritm for speaker change detection," in *Eurospeech'03*, 2003.

[43] M. Roch and Y. Cheng, "Speaker segmentation using the map-adapted bayesian information criterion," in *Odissey-04*, Toledo, Spain, 2004, pp. 349–354.

[44] P. Delacourt, D. Kryze, and C. J. Wellekens, "Speaker-based segmentation for audio data indexing," in *ESCA Workshop on accessing Information in Audio Data*, 1999.

[45] B. Zhou and J. H. Hansen, "Unsupervised audio stream segmentation and clustering via the bayesian information criterion," in *ICSLP-2000*, vol. 3, Beijing, China, 2000, pp. 714–717.

[46] J.-F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, and C. Wellekens, "A speaker tracking system based on speaker turn detection for nist evaluation," in *ICASSP-00*, Istanbul, Turkey, 2000, pp. 1177–1180.

[47] H. Gish, M.-H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *ICASSP-91*, vol. 2, Toronto, Canada, 1991, pp. 873–876.

[48] G. Gish and M. Schmidt, "Text-independent speaker identification," *Signal Processing Magazine, IEEE*, pp. 18–32, October 1994.

[49] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *DARPA Speech Recognition Workshop*, Chantilly, 1997, pp. 97–99.

[50] J. Hung, H. Wang, and L. Lee, "Automatic metric based speech segmentation for broadcast news via principal component analysis," in *ICSLP'00*, Beijing, China, 2004.

[51] X. Anguera and J. Hernando, "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, november 2004.

[52] T. Wu, L. Lu, K. Chen, and H.-J. Zhang, "Ubm-based real-time speaker segmentation for broadcasting news," in *ICASSP'03*, 2003.

[53] ——, "Ubm-based incremental speaker adaptation," in *ICME'03*, vol. 2, 2003, pp. 721–724.

[54] ——, "Universal background models for real-time speaker change detection," in *International Conference on Multimedia Modeling*, 2003.

[55] L. Lu and H.-J. Zhang, "Speaker change detection and tracking in real-time news broadcasting analysis," in *ACM International Conference on Multimedia*, 2002, pp. 602–610.

[56] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, October 2002.

[57] H. S. Beigi and S. H. Maes, "Speaker, channel and environment change detection," in *World Congress on Automation*, 1998.

[58] A. G. Adami, S. S. Kajarekar, and H. Hermansky, "A new speaker change detection method for two-speaker segmentation," in *ICASSP'02*, Orlando, Florida, 2002.

[59] S. Nakagawa and H. Suzuki, "A new speech recognition method based on vq-distorsion and hmm," in *ICASSP-93*, vol. 2, Minneapolis, USA, April 1993, pp. 676–679.

[60] D. Liu and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *Eurospeech-99*, vol. 3, Budapest, Hungary, 1999, pp. 1031–1034.

[61] H. Meinedo and J. Neto, "Audio segmentation, classification and clustering in a broadcast news task," in *ICASSP'03*, Hong-Kong, China, 2003.

[62] H. Jin, F. Kubala, and R. Schwartz, "Automatic speaker clustering," in *DARPA Speech Recognition workshop*, Chantilly, USA, 1997.

[63] A. Solomonov, A. Mielke, M. Schmidt, and M. Gish, "Clustering speakers by their voices," in *ICASSP'98*, vol. 2, Seattle, USA, 1998, pp. 757–760.

[64] W.-H. Tsai, S.-S. Cheng, and H.-M. Wang, "Speaker clustering of speech utterances using a voice characteristic reference space," in *ICSLP-04*, Jeju Island, Korea, 2004.

[65] A. Sankar, F. Beaufays, and V. Digalakis, "Training data clustering for improved speech recognition," in *Eurospeech-95*, Madrid, Spain, 1995.

[66] L. Heck and A. Sankar, "Acoustic clustering and adaptation for robust speech recognition," in *Eurospeech-97*, Rhodes, Greece, 1997.

[67] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system," in *Rich Transcription Workshop*, New Jersey, USA, 2004.

[68] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, "Segmentation od speech using speaker identification," in *ICASSP-94*, vol. 1, Adelaide, Australia, April 1994, pp. 161–164.

[69] Y. Moh, P. Nguyen, and J.-C. Junqua, "Towards domain independent speaker clustering," in *ICASSP-03*, Hong Kong, 2003.

[70] S. Johnson and P. Woodland, "Speaker clustering using direct maximization of the mllr-adapted likelihood," in *ICSLP-98*, vol. 5, 1998, pp. 1775–1779.

[71] S. Johnson, "Who spoke when? - automatic segmentation and clustering for determining speaker turns," in *Eurospeech-99*, Budapest, Hungary, September 1999.

[72] F. Bimbot and L. Mathan, "Text-free speaker recognition using an arithmetic-harmonic sphericity measure," in *Eurospeech'93*, Berlin, Germany, 1993, pp. 169–172.

[73] Nist spring rich transcription evaluation in meetings website. [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2005/spring

[74] S. Cassidy, "The macquarie speaker diarization system for rt04s," in *NIST 2004 Spring Rich Transcrition Evaluation Workshop*, Montreal, Canada, 2004.

[75] C. Fredouille, D. Moraru, S. Meignier, L. Besacier, and J.-F. Bonastre, "The nist 2004 spring rich transcription evaluation: Two-axis merging strategy in the context of multiple distant microphone based meeting speaker segmentation," in *NIST 2004 Spring Rich Transcrition Evaluation Workshop*, Montreal, Canada, 2004.

[76] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre, "The elisa consortium approaches in broadcast news speaker segmentation during the nist 2003 rich transcription evaluation," in *ICASSP-04*, Montreal, Canada, 2004.

[77] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel, "Speaker segmentation and clustering in meetings," in *NIST 2004 Spring Rich Transcrition Evaluation Workshop*, Montreal, Canada, 2004.

[78] Icsi meetings recorder corpus. [Online]. Available: http://www.icsi.berkeley.edu/Speech/mr

[79] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," in *ICCASP-03*, Hong Kong, 2003.

[80] Cmu meetings corpus website. [Online]. Available: http://penance.is.cs.cmu.edu/meeting_room

[81] S. Burger, V. Maclaren, and H. Yu, "The isl meeting corpus: The impact of meeting type on speech style," in *ICSLP-02*, Denver, USA, 2002.

[82] Nist pilot meeting corpus website. [Online]. Available: http://www.nist.gov/speech/test_beds/mr_proj/meeting_corpus_1

[83] National institute for standards and technology. [Online]. Available: http://www.nist.gov/speech

[84] Nist fall rich transcription evaluation in broadcast news website. [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2004/fall