A MULTIMODAL APPROACH TO INITIALISATION FOR TOP-DOWN SPEAKER DIARIZATION OF TELEVISION SHOWS

Simon Bozonnet[1], *Félicien Vallet*[2,3], *Nicholas Evans*[1], *Slim Essid*[2], *Gaël Richard*[2] *and Jean Carrive*[3]

[1]EURECOM, BP193, F-06904 Sophia Antipolis Cedex, FRANCE
[2]LTCI, TELECOM ParisTech and CNRS, 46 rue Barrault, 75634 Paris cedex 13, FRANCE
[3]INA, 4 avenue de l´ Europe, 94366 Bry-sur-Marne cedex, FRANCE
emails: {lastname}@eurecom.fr, {firstname.lastname}@telecom-paristech.fr, jcarrive@ina.fr

ABSTRACT

This paper presents a new multimodal approach to speaker diarization of TV show data. We hypothesize that the intra-speaker variation in visual information might be less than that in the corresponding acoustic information and therefore might be better suited to the task of speaker model initialisation. This is an acknowledged weakness of the computationally efficient top-down approach to speaker diarization that is used here. Experimental results show that a recently proposed approach to purification and the new multimodal approach to initialisation together deliver 22% and 17% relative improvements in diarization performance over the baseline system on independent development and evaluation datasets respectively.

1. INTRODUCTION

Speaker diarization is now a main-stream speech processing research topic and involves determining the number of speakers in an audio document and the intervals when each speaker is active, a task otherwise referred to as 'who spoke when?' Among other previously popular domains of telephone conversations and broadcast news, it is today that of conference meetings which is widely considered to be the most challenging and accordingly attracts the most attention. Conference meetings are also the focus of the internationally competitive NIST Rich Transcription (RT) evaluations [1]. Among other specific attributes, the highly spontaneous nature of meetings pose several challenges to speaker diarization systems, many of which remain problematic, e.g. the detection of overlapping speech and effective system combination strategies.

Since the focus on conference meeting data has somewhat of a narrow application domain, researchers are already looking to new opportunities. Speaker diarization has utility in any application where multiple speakers may be expected and, with the mass of multimedia information now available, it is arguably for speaker indexing and content structuring that speaker diarization has the greatest potential.

In recent months we have started some activities in speaker diarization for mainstream multimedia data and, due to the immediate availability of the 'Grand Échiquier' database[1], we have thus far focused our efforts on broadcast television (TV) talk-shows. The application of speaker diarization to new domains is notoriously troublesome and it is common for systems that are optimised on one domain to perform poorly when applied without modification to different data. This recent experience has proved no different and the performance of our baseline system, that is optimised for conference meeting data, performs poorly when applied to TV show data.

The TV show data considered here contains far more speakers than do typical conference meetings, a greater spread of speaker floor time and more rapid speaker turns. It can thus prove quite difficult to detect speakers and therefore to initialise speaker models. Initialisation is a well known weakness of top-down approaches to speaker diarization; [2, 3] bottom-up approaches are arguably better suited to this particular task. However, the top-down approach is particularly computationally efficient and it is therefore of interest to improve its robustness for large scale applications such as indexing and content structuring.

A large volume of data in such tasks is multimodal yet traditional approaches to speaker diarization exploit only acoustic information. Some earlier work investigated the utilisation of visual information for speaker diarization but most of it focuses on conference meeting data, e.g. [4], which utilised a standard bottom-up approach to speaker diarization and [5], which used BIC-based segmentation and graph spectral partitioning for clustering. To our knowledge, none of the existing work has involved top-down approaches. This paper therefore reports the first attempt to utilise visual information to improve performance in a top-down approach to speaker diarization for large scale multimedia tasks. Due to the weaknesses of top-down approaches, in this first attempt, we concentrate on utilising visual features only at the initialisation level.

The remainder of this paper is organised as follows. Section 2 gives brief details of the multimodal database on which we report experimental results and discusses the differences between it and typical conference meeting recordings. Section 3 describes our baseline diarization system and the modifications which were necessary in order to apply it successfully to the new database. Section 4 describes how visual features are utilised and our experimental work to assess their benefit is reported in Section 5. Finally, our conclusions are presented in Section 6.

2. TV SHOWS VS MEETINGS

The baseline speaker diarization system used in this work was developed for the conference meeting domain, which is the focus of current NIST RT evaluations. In this paper we report experiments on a corpus comprised of over 50 French-language, 'Grand Échiquier' (GE) TV talk-show programmes from the 1970-80s. Each show focuses on a main guest, and other supporting guests, who are both interviewed

[1]Distributed by the French Institut National de l'Audiovisuel (INA) www.ina-sup.com/en/

by a host presenter. The interviews are punctuated with film excerpts, live music and other performances. The database presents numerous characteristics and challenges that have made it popular among both national and European multimedia research projects, e.g. the European K-Space network of excellence [6].

The speaker diarization of such data is especially challenging and there are numerous differences between conference meetings and TV shows. Among the most obvious are those related to recording quality. Meetings are generally recorded using distant wall-mounted or desktop microphones. The distances between speakers and microphones can vary greatly and may change throughout the recording if speakers turn their heads or move around the meeting room. In contrast, TV shows are usually recorded with high-quality boom and/or lapel microphones and therefore the signal-to-noise ratio is often much better than it is for meeting recordings.

The better audio quality of TV shows should be to our advantage. However, perhaps surprisingly, and as we explain later, speech activity detection tends to be more challenging for TV shows than it is for meetings. In TV shows, aside from the presence of film excerpts, live music, audience applause and laughter, silences during speaker turns can be very short or almost negligible. Compared to meetings, where speakers often pause to collect their thoughts or to reflect before responding to a question, TV show speech tends to be more fluent and sometimes almost scripted. This is perhaps due to the fact that the main themes and discussions are prepared in advance and known by the speakers.

Quantitative differences between TV shows and conference meetings are summarised in Table 1 which illustrates various statistics (column 1) for 7 TV shows (column 2) from the GE database, which have thus far been annotated according to standard NIST RT protocols [1], and the 7 conference meetings from the NIST RT'09 dataset (column 3). The average show length for the GE and RT'09 dataset is 147 minutes and 25 minutes respectively. On average there are 50 minutes (GE) and 13 minutes (RT'09) of speech per show (i.e. with noise and music removed). In the GE dataset there are an average of 1033 speech segments per show with an average length of 3 seconds (cf. 882 segments with an average length of 2 seconds for the RT'09 dataset). There are also differences in the amount of overlapping speech (averages of 5 minutes cf. 3 minutes per show). As a fraction of the average speech time the percentage of overlapping speech in each case is 10% (GE) and 23% (RT'09) and thus there is less overlapping speech in the GE dataset than there is in the RT'09 dataset.

Finally, we consider differences in speaker statistics. Also illustrated in Table 1 are the average number of speakers and the average floor time for the most and least active speakers in each show. On average there are 13 speakers per TV show and only 5 speakers per conference meeting. This might be expected given the longer average length of TV shows. Given a larger number of speakers we can expect a smaller average inter-speaker difference than for meetings and hence increased difficulties in speaker diarization. Furthermore, we see that the spread in floor time is much greater for the GE dataset than it is for the RT'09 dataset. The average speaking time for the most active speaker is 1476 seconds for the GE dataset (cf. 535 seconds for RT'09) and corresponds to the host presenter in each case. The average speaking time for the least active speaker is only 7 seconds (cf. 146 seconds for RT'09) and corresponds to one of the minor supporting guests. Speakers with such little data are extremely difficult to detect and thus this aspect of the TV show dataset is likely to pose significant difficulties for speaker diarization even if, according to standard NIST protocols, each speaker's contribution to the diarization performance metric is time weighted. Furthermore, the presence of one or two dominant speakers means that lesser active speakers will be comparatively harder to detect, even if they too have a significant floor time.

Attribute	GE	NIST RT'09
No. of shows	7	7
Evaluation time	147 min.	25 min.
Total speech	50 min.	13 min.
No. of segments	1033	882
Av. segment length	3 sec.	2 sec.
Overlap	5 min.	3 min.
No. speakers	13	5
most active	1476 sec.	535 sec.
least active	7 sec.	146 sec.

Table 1: A comparison of Grand Échiquier (GE) and NIST RT'09 database characteristics.

Even if there is less overlapping speech the nature of TV shows thus presents unique challenges not seen in meeting data: the presence of music and other background non-speech sounds, shorter pauses, a greater spread in speaker floor time and more speakers. These issues are likely to exacerbate weaknesses with initialisation and thus we seek to improve performance by utilising video features.

3. SYSTEM DESCRIPTION

In this section we describe a baseline top-down speaker diarization system and then some modifications which are necessary so that it may be applied successfully to TV show data. The system described here is audio-only. A multimodal approach is described later in Section 4.

The baseline diarization system adopted here is that of LIA-EURECOM's submission [2] to the NIST RT'09 evaluation [1]. Developed by LIA, the system is based upon an evolutive hidden Markov model (E-HMM) [7] approach to speaker diarization where states correspond to speakers and transitions between states correspond to speaker turns. Speakers are modelled with Gaussian mixture models (GMMs). A full description of the system is available in [2] and accordingly only a brief system summary is reported here. The system is composed of four stages, each one of which is summarised below with a recently introduced purification stage [3].

Speech activity detection (SAD) is the first step and is performed by alignment to a 2-state HMM with speech and non-speech models. Several iterations of decoding and adaptation are performed and produce the speech/non-speech labels which are used in subsequent stages.

Segmentation and clustering aims to identify the speakers and when each of them is active. First, a general GMM model is fitted to all the speech available in the recording with an expectation maximisation (EM) algorithm. A new speaker is then identified with the selection of single segment

currently assigned to the general GMM and a new speaker model is trained, again with EM. Several iterations of Viterbi decoding and adaptation are performed to give a new segmentation hypothesis. New speakers are added one-by-one, in identical fashion, and the process stops when there remains no more eligible segments for model initialisation.

Purification was added recently [3] and was inspired by the segmental initialisation approach proposed in [8]. The aim is to purify the clusters by retraining new speaker models using only the sub-segments which best fit each model and by reassigning the other sub-segments to the nearest new model via several iterations of Viterbi decoding and adaptation.

Resegmentation is applied to refine the speaker boundaries and to delete irrelevant speakers (speakers with too little speech). In contrast to the previous segmentation and clustering step the models are incorporated simultaneously into the HMM, and the models are tuned through the MAP adaptation of a world model that is trained on external data.

Normalization and resegmentation involves a final pass of resegmentation but on feature vectors that are normalised segment-by-segment to fit a zero-mean and unity-variance distribution. Full details are available in [2].

The baseline system was developed for conference meeting data and our preliminary attempts to apply the same system to TV show data produced poor results. Some minor modifications were necessary so that the system can be applied successfully to TV show data.

Non-speech periods in the TV show data are mainly music, applause or laughter. Since we have not implemented a music detector we assume that the few music intervals are known and thus they are manually removed. Also, as described above, speech pauses are far less common than they are in meeting data. Since the penalty incurred by ignoring speech pauses is greater than that incurred by trying to detect them (i.e. it leads to high levels of missed speech), and since there is in any case very few genuine non-speech intervals, we decided to skip the SAD step for TV show data.

The recently introduced purification step was also optimized for meetings and did not give good performance when applied to the TV show data. This is mainly due to inactive speakers for which, after purification, there remain insufficient data with which to retrain new speaker models. The approach still delivers improved performance for speakers with sufficient data and so purification is here only applied to speakers who, following segmentation and clustering, are deemed to be active for more than 14 seconds.

Finally, the normalization step, whose purpose for meeting data includes channel compensation to reduce the effects of differing distances between microphones and speakers, was found not to bring any consistent performance improvement for the TV show data, where recordings are made in far more controlled and consistent conditions. This step is therefore also skipped.

It is acknowledged that the manual labelling of music intervals renders our experiments artificial. However, it is reasonable to assume that automatic music detection errors should have equivalent effects on speaker diarization system performance both with and without visual features and so it should not detract too significantly from the assessment reported here. Further more, even though we do not make any effort to detect non-speech intervals they are nonetheless included for scoring purposes, as dictated by standard NIST speaker diarization assessment protocols.

4. MULTIMODAL APPROACH

TV show data is edited; shot selection is performed by a TV director who generally tries to focus on the active speaker. Therefore we can assume that, most of the time, '*we see who we hear*'. However, the task is to determine who is *speaking*, not who we *see*, and thus it is still the acoustic signal that carries the most pertinent information. Since it is not necessarily the case that acoustic and visual information are correlated in terms of speakers, multimodal feature combination or fusion can be problematic in speaker diarization tasks and so standard approaches to combination or fusion are not appropriate. For this reason, and due to the initialisation weaknesses of top-down speaker diarization systems, we thus propose to use visual features as early as possible in the process and here consider their use only for initialization.

Our hypothesis is that, even though visual features might not always reflect the active speaker, for an unsupervised task such as speaker diarization, they are better suited to initialisation than are acoustic features since they are more stable and consistent, i.e. whereas the acoustic content will surely change, certain aspects of a speaker's appearance, namely their clothing, will surely not.

The idea is to perform unsupervised pre-clustering with visual features to over-cluster the data into a pool of small pre-clusters whose number should exceed the true number of speakers. A candidate cluster is then selected, according to some criteria, and is used to introduce a new speaker into the E-HMM. This is done using the corresponding acoustic features in exactly the same manner as before. New speakers are added one-by-one, but now using the pre-clusters for initialisation, and the process is repeated until there are no more remaining candidate clusters. Except for the model initialisation stage the system is identical to that described in Section 3. In the following we describe our choice of visual features and the approach to pre-clustering.

4.1 Visual features

On a TV set clothing is often carefully chosen so that participants are easily distinguishable and to avoid clothing clashes. Therefore we expect that features which characterise faces or clothing should be of use for speaker diarization.

Face detection is performed according to the popular Viola and Jones method [9] with the software available in the OpenCV library [10]. From identified faces bounding boxes are then determined according to a scaled rectangle situated immediately below the face, similar to the method described in [11]. An example is illustrated in Figure 1 where the green and red rectangles show the bounding boxes for faces and clothing respectively. Colour features are then extracted from the clothing region.

A total of 22 visual features were considered (not reported here) and were ranked according to their speaker discriminability according to the method used in [12]. This analysis showed that the feature based on the average dominant clothing colour had the highest speaker discriminability and is that used for visual pre-clustering experiments reported here.

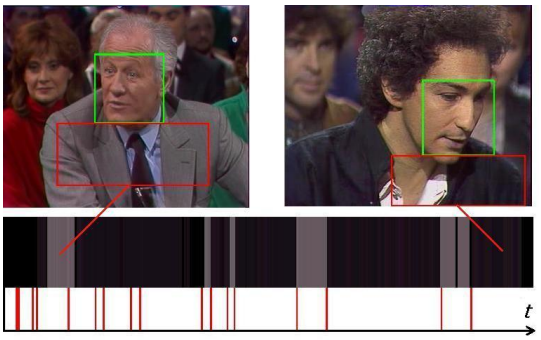

Figure 1: An illustration of face (green) and clothing (red) bounding boxes. The dominant, on-screen clothing colour is illustrated below against time and corresponds to the active speaker. Changes in the dominant clothing colour can indicate a speaker turn.

4.2 Pre-clustering

Since the guest lineup often changes between musical intervals we first segment the show into non-musical (i.e. speech) intervals and treat each individually. We suppose that each speech segment contains between 2 and 10 speakers and we apply a classical k-means clustering [13] to the average dominant clothing colour feature to partition the visual observations into a number of clusters. So as to reduce the chances of a single cluster attracting data from more than a single speaker we aim to identify more clusters than there are speakers and have adopted the method proposed by Sugar and James in [14] which determines the appropriate number of clusters automatically. We only keep clusters with more than 10 seconds of assigned observations. Clusters with fewer than 10 seconds of observations are removed and their data is reassigned to other clusters. The procedure usually results in more than one model per speaker.

Then, new models are trained using the acoustic data which corresponds to each of the pre-clusters. The new models are then purified using the approach described in [3] but, to accommodate an increased spread in speaker floor time, we use a purity factor of 75% (cf. 55% in [3]). This produces a new set of pre-clusters.

Our experiments have shown that it is important to add the most dominant speaker to the E-HMM ahead of less dominant speakers and so the five pre-clusters which are assigned the most data are selected as potential candidates for adding the first speaker model into the E-HMM. Assuming that pre-clusters of good quality (i.e. those which largely correspond to a single speaker) will attract frames from fewer, but concentrated segments, rather than a large number of short, fragmented segments, we then compute the ratio of the total amount of data and the number of segments in each of the five pre-clusters. The pre-cluster with the highest ratio of frames-to-segments is selected as the first speaker and a new model is thus added to the E-HMM, which is updated in the usual way. The four other pre-clusters are moved back into the pool of pre-clusters which are used to add subsequent speakers.

Additional speakers are added to the E-HMM by choosing the next pre-cluster which has the largest amount of data currently assigned to the general GMM model in the E-HMM. Speaker models are added one-by-one, in the same way as before, until there remain no more pre-clusters with more than 6 seconds of data assigned to the general GMM.

System	Dataset	SpkError	SAD	DER
Baseline System	GE dev.	25.7/26.0	15.2/9.7	40.8/35.6
Baseline + Pur.	GE dev.	23.6/23.4	15.2/9.7	38.7/33.0
Optimized audio system + Pur.	GE dev.	21.9/21.5	11.8/5.8	33.6/27.3
Multimodal system	GE dev.	22.0/23.8	11.8/5.8	33.8/29.6
Multimodal system + Pur.	GE dev.	18.3/19.5	11.8/5.8	30.0/25.3
Baseline System	GE eval.	30.4/31.1	9.4/5.5	39.7/36.5
Baseline + Pur.	GE eval.	28.8/29.5	9.4/5.5	38.2/34.9
Optimized audio system + Pur.	GE eval.	22.9/25.9	7.4/3.4	30.3/29.3
Multimodal system	GE eval.	24.9/26.2	7.4/3.4	32.3/29.6
Multimodal system + Pur.	GE eval.	24.2/25.5	7.4/3.4	31.6/28.8
Baseline	RT'09	17.6/18.3	8.4/3.2	26.0/21.5
Baseline + Pur.	RT'09	12.7/12.8	8.4/3.2	21.1/16.0

Table 2: Speaker diarization performance on the GE dataset (development and evaluation subsets) and the NIST RT'09 dataset with different system configurations. Illustrated are the contributions of speaker error (SpkError) and speech activity detection (SAD) performance to the total combined diarization error rate (DER). In all cases error rates are given with/without scoring overlapping speech regions.

5. EXPERIMENTAL WORK

Our experimental results are summarised in Table 2 and aim to demonstrate the potential merit of a multimodal approach to speaker diarization. We report experiments on two subsets of the GE database. The 7 annotated shows are divided into a development set of 4 shows and an evaluation set of 3 shows. We acknowledge that the number of shows, and therefore the statistical significance, is small. However, each of the GE shows is recorded in almost identical conditions and therefore the average inter-show difference is likely to be less than it is for a typical NIST RT dataset. Conference meetings are recorded at different sites, using different acquisition equipment and different room layouts etc. The TV studio is, in contrast, mostly the same. Table 1 shows that there is an average of only 13 minutes of speech per show in the RT'09 dataset which amounts to a total of 91 minutes of speech for the whole dataset. The GE evaluation set of 3 shows has an average of 50 minutes of speech per show. This amounts to a total of 150 minutes of speech. Therefore we have less inter-show variation and considerably more speech than there is in a standard NIST RT speaker diarization dataset.

To facilitate the comparison of performance to the work of others, all results in Table 2 are presented with/without the scoring of overlapping speech. In the following, unless explicitly stated otherwise, we only discuss scores which include the scoring of overlapping speech. The total diarization error (DER) is illustrated with the contributions from speaker errors (SpkErr) and speech activity detection (SAD). The second line of Table 2 shows performance when the baseline speaker diarization system of [2] was applied directly to the GE dataset without modification. This system is that described in the start of Section 3 but does not include the recently introduced purification stage. A total diarization error rate of 40.8% corresponds to an SpkErr of 25.7% and SAD errors of 15.2%. For comparison we illustrate in line 12 of Table 2 the performance obtained when the exact same system is applied to the NIST RT'09 database. Results here are identical to those published in [2]. For the meeting data, corresponding results are a total DER of 26.0% (17.6% SpkErr and 8.4% SAD). The 3rd and 13th lines of Table 2 show

performance on the GE and RT'09 datasets with integrated purification, as described in the start of Section 3. This system corresponds to that published in [3]. Respective DERs of 38.7% and 21.1% show a consistent improvement across the two datasets.

Performance is considerably worse for TV shows than it is for meetings. Both SpkErr and SAD performance are poor for the TV show data (23.6 % and 15.2% respectively cf. 12.7 % and 8.4% for meetings) but significant improvements in performance are obtained with the system modifications proposed toward the end of Section 3, namely those of removing SAD, optimized purification and no normalisation. Corresponding results are illustrated on line 4 which show a total DER of 33.6%. There are improvements in both SpkErr and SAD error rates (21.9% and 11.8% respectively). Over the baseline system with purification (line 3) this corresponds to a relative improvement in DER of 13%. The SpkErr remains high, however, and is caused by the poor detection of relatively inactive speakers..

When we perform initialisation with visual features, according to the system described in Section 4, but without purification, we obtain a total DER of 33.8% (line 5). Thus similar levels of improvement are obtained with purification and visual features. When we combine purification and initialisation with the use of visual features we obtain an average DER of 30.0% (line 6). Therefore the recently introduced purification module, and the approach to initialisation with visual features that is proposed here, bring complementary improvements to speaker diarization performance. Compared to the baseline system with purification (line 3) this corresponds to a relative improvement of 22% in DER and is attributed to improvements in speaker model purity and the better detection of relatively inactive speakers.

All of the above results correspond to systems that are optimised for the development set. To validate our findings on unseen data we repeated the experiments on the evaluation set and observed a similar trend in performance. The original baseline system without purification gives an average DER of 39.7% (line 7). With purification performance improves to 38.2% (line 8). Without SAD, optimised purification and no normalisation, we obtain 30.3% (line 9). Using visual features for initialisation, but no purification, we obtain 32.3% (line 10). Finally, when we combine purification and visual features we obtain a DER of 31.6% (line 11). These results are marginally worse than the results for the optimised audio system with purification (line 9) but do not discount the merit of visual features. These scores include overlapping speech even though we do not attempt to detect overlap. We note that when these regions are not scored, we achieve a small gain in performance with visual features and purification (29.3% cf. 28.8%). Referring once again to scores including overlapping speech, this corresponds to a relative improvement of over 17% compared to our baseline system with purification (line 8). The combined approaches thus deliver complementary improvements in DER on both development and evaluation datasets and serve to both validate the efficiency of our purification step introduced in [3] and the merit of video features for initialisation.

6. CONCLUSION

This paper reports our first attempts to utilise visual information to assist with speaker diarization. Experiments are reported on a dataset of 7 TV shows. Whilst the two development and evaluation subsets contain fewer files than a typical NIST RT dataset, they both contain more speech and should have less inter-show variation.

Based on the hypothesis that we often '*see who we hear*', our assumption that visual features are better suited to initialisation than are acoustic features and due to the acknowledged weaknesses of the computationally efficient top-down approach to speaker diarization, we investigate the use of visual information for initialisation purposes only. Experimental results show that whilst diarization performance is lower than that reported for conference meeting data, a recently proposed purification step and the use of visual features give complementary improvements in speaker diarization performance and relative improvements in DER of 22% and 17% on the development and evaluation sets respectively.

The paper thus establishes the potential of visual information for intialisation purposes, in particular for the identification of relatively inactive speakers. With such a computationally efficient, top-down approach to speaker diarization there is thus potential for large scale indexing and content structuring applications. This work is however a first attempt and future work should focus on strengthening integration of visual features.

7. ACKNOWLEDGEMENTS

The work reported in this paper was partly funded by the French Institut-Télécom SELIA project.

REFERENCES

[1] NIST, "The NIST Rich Transcription 2009 (RT'09) evaluation," http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf, 2009.

[2] C. Fredouille, S. Bozonnet, and N. Evans, "The LIA-EURECOM RT09 Speaker Diarization System," in *RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA*, 2009.

[3] S. Bozonnet, N. W. D. Evans, and C. Fredouille, "THE LIA-EURECOM RT'09 speaker diarization system : enhancements in speaker modelling and cluster purification," in *ICASSP 2010, to appear*, March 2010.

[4] G. Friedland, H. Hung, and C. Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *International Conference on Acoustics, Speech and Signal Processing*, 2009.

[5] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi, "Audio segmentation and speaker localization in meeting videos," in *International Conference on Pattern Recognition*, 2006.

[6] K-Space, "The European K-Space Network Of Excellence," http://www.k-space.eu/.

[7] S. Meignier, J.F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2001, pp. 175–180.

[8] T. Nguyen et al., "The IIR-NTU Speaker Diarization Systems for RT 2009," in *RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA*, 2009.

[9] P. Viola and M. Jones, "Robust real-time object detection," in *International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling*, 2001.

[10] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*, O'Reilly Media, 2008.

[11] G. Jaffr and P. Joly, "Costume: A new feature for automatic video content indexing," in *International conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*, 2004.

[12] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, "Prosodic and other long-term features for speaker diarization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 5, pp. 985–993, July 2009.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, 2000.

[14] C. A. Sugar and G. M. James, "Finding the number of clusters in a data set : An information theoretic approach," *Journal of the American Statistical Association*, vol. 98, pp. 397–408, 2003.