On the Applicability of Speaker Diarization to Audio Concept Detection for Multimedia Retrieval

Robert Mertens
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
rmertens@icsi.berkeley.edu

Po-Sen Huang
Beckman Institute, ECE Department
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
huang146@illinois.edu

Luke Gottlieb
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
luke@icsi.berkeley.edu

Gerald Friedland
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
fractor@icsi.berkeley.edu

Ajay Divakaran
SRI International Sarnoff
201 Washington Road
Princeton, NJ 08540, USA
ajay.divakaran@sri.com

***Abstract*—Recently, audio concepts emerged as a useful building block in multimodal video retrieval systems. Information like "this file contains laughter", "this file contains engine sounds" or "this file contains slow music" can significantly improve purely visual based retrieval. The weak point of current approaches to audio concept detection is that they heavily rely on human annotators. In most approaches, audio material is manually inspected to identify relevant concepts. Then instances that contain examples of relevant concepts are selected – again manually – and used to train concept detectors. This approach comes with two major disadvantages: (1) it leads to rather abstract audio concepts that hardly cover the audio domain at hand and (2) the way human annotators identify audio concepts likely differs from the way a computer algorithm clusters audio data – introducing additional noise in training data. This paper explores whether unsupervized audio segementation systems can be used to identify useful audio concepts by analyzing training data automatically and whether these audio concepts can be used for multimedia document classification and retrieval. A modified version of the ICSI (International Computer Science Institute) speaker diarization system finds segments in an audio track that have similar perceptual properties and groups these segments. This article provides an in-depth analysis on the statistic properties of similar acoustic segments identified by the diarization system in a predefined document set and the theoretical fitness of this approach to discern one document class from another.**

***Keywords*-Audio Clustering, Audio Indexing, Speaker Diarization, Video Indexing**

I. Introduction

Multimedia retrieval becomes more and more important due to a number of reasons. The amount of multimedia data posted by end users on the web is increasing on a daily basis. Surveillance data is gathered with unprecedented coverage and archives of professionally created entertainment media or documentaries are growing steadily. All of these media objects are, however, of little value if users can not find them and retrieve them. At this point structuring and indexing of this vast amount of multimedia data makes the difference. In order to tackle the challenge of indexing multimedia data, a multitude of approaches have been devised in the past (see [1] or [2] for an overview). A video usually contains and audio and a visual stream, however many approaches for video analysis focus only on the visual part of a video. Audio has recently begun to play a role in multimodal media analysis and can be leveraged to complement results from visual analysis to increase the effectiveness of multimedia retrieval or detection approaches. Audio information can be used to these ends in two fundamentally different ways. Speech recognition has been employed for video analysis since the late 1990s [3]. The second method for using audio analysis for video analysis is the detection of sound concepts that describe a video's content. The presence of human defined lower level acoustic concepts such as "indoor sound" or "people laughing" conveys valuable information as to a video's content and such sound concepts can be automatically detected once a system is trained to recognize them [4][5]. The use of low level acoustic concepts does, however usually involve manual concept definition. The two downsides of manual concept definition are that it usually leads to rather abstract concepts and that it introduces a human bias as human annotators are likely to identify these concepts based on different properties of sound than a computer algorithm would. This paper explores the applicability of a speaker diarization engine to the definition and extraction of low level acoustic concepts from domain specific training data. The speaker diarization engine clusters segments of an audio stream that exhibit similar properties. It thus extracts acoustic concepts as they are defined by a computer algorithm - namely the speaker diarization engine. To explore whether this assumption holds, we have

978-0-7695-4589-9/11 $26.00 © 2011 IEEE
DOI 10.1109/ISM.2011.79

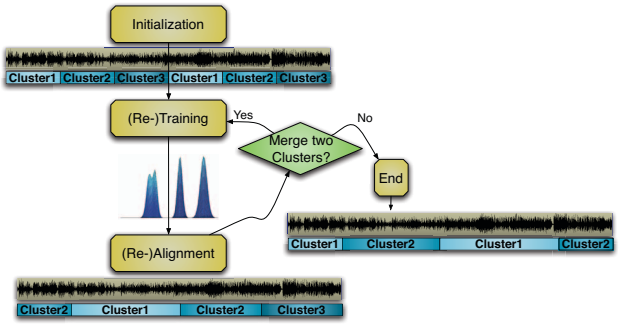


Figure 1. ICSI Speaker Diarization System

generated and examined diarization data from the TRECVID MED 2011 data set. This data set has been released by NIST as training data for the TRECVID MED 2011 concept detection challenge. It contains randomly selected videos that are examples for fifteen different categories of high level concepts such as ”wedding ceremony” and ”woodworking project” and thus represents a useful data set for the analysis presented in this paper. The data set does not only deliver a wealth of low level features that can be detected by the diarization approach, it also separated data into higher level classes so that we can explore whether higher level classes can be predicted by the absence or presence of certain low level features. The analysis of the distribution of speaker segments found in the videos from different categories shows that speaker segments are not randomly distributed but can be used to predict whether a video belongs to a certain event class or not. It thus indicates that speaker segmentation generates low level audio concepts that can be used for higher level machine learning based classification. The remainder of this paper is organized as follows: Section 2 briefly explains the ICSI speaker diarization system. The contents of the NIST TRECVID 2011 MED data set are discussed in section 3. Section 4 presents the methodology of the experiment. An overview of significant low level features found in the experimental results as well as a discussion of the distribution of these results is given in Section 5. Section 6 discusses the relevance of the findings of this paper and touches upon the perspectives opened for future research.

II. ICSI Speaker Diarization System

For detecting sound concepts in each individual video we used a system based on the ICSI speaker diarization system [7] in a faster-than-realtime version [9]. The actual diarization process consists of a pre-processing phase and a segmentation and clustering phase, as shown in figure 1. In the preprocessing phase audio features, in this case Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from the video soundtrack. We use a frame period of 10 ms with an analysis window of 30 ms in the feature extraction. In its original application context the speaker diarization system also employs speech/non-speech segmentation to exclude non-speech in later processing steps. In order to use all audio information, we have omitted the exclusion of non-speech segments. In the segmentation and clustering stage of speaker diarization, an initial segmentation is generated by uniformly partitioning the audio track into K segments of the same length. K is chosen to be much larger than the assumed number of speakers in the audio track. For meeting recordings of about 30 minute length, previous work [6] experimentally determined K = 16 a good value. For the audio tracks used in the TRECVID MED 2011 data set, we have determined K=64 a suitable value. The main reason for the higher K value is that the number of significant audio concepts in a video is much higher than the average number of speakers in a meeting video. The procedure for diarization is shown in Figure 1 and takes the following steps, a more detailed description can be found in [7]:

1) Initialization: Train a set of Gaussian Mixture Models (GMMs), one for each initial cluster.

2) Re-segmentation: Re-segment the audio track using the current GMMs using majority vote on the likelihoods of a specified minimum duration [8]. For audio concept detection, we have set this minimum duration to 200 milliseconds in order to capture sounds of smaller duration. For speaker segmentation higher values are used. A typical mimimum for speaker segmentation would be 2500 milliseconds

3) Re-training: Retrain the GMMs on current segmentation using the expectation-maximization (EM) algorithm [8].

4) Agglomeration: Select the closest pair of clusters and merge them. At each iteration, the algorithm checks all possible pairs of clusters to see if there is an improvement in BIC scores by merging each pair and re-training it on the combined audio segments. The clusters from the pair with the largest improvement in Bayesian Information Criterion (BIC) scores are merged and the new GMM is used. The algorithm then repeats from the re-segmentation step until there are no remaining pairs that will lead to an improved BIC score.

The result of the algorithm consists of a segmentation of the audio track with n clusters and with one GMM for each cluster, where n is assumed to be the number of speakers, which in our case are audio concepts.

III. TRECVID MED 2011 DEV-T Data Set

The TRECVid 2011 MED dataset is different from the original TRECVid dataset. The MED dataset is comprised of “found videos”, i.e. consumer-produced videos downloaded from various social networking sites. Most videos are very short (a couple of minutes) and not produced professionally. The query sets (so-called event kits) are comprised of fifteen categories with only five of those categories available in the testing set. The event kits consist of a total of 2040 videos and the test set of a total of 4251 videos. The five event categories which are available in the test set are “attempting

Table I
NUMBER OF VIDEOS FOR TRAIN AND TEST

Category	Description	Train Data	Test Data
E001	Board Tricks	160	111
E002	Feeding Animal	160	111
E003	Landing Fish	122	86
E004	Wedding	128	88
E005	Woodworking	142	100
E006	Birthday Party	173	0
E007	Changing Tire	110	0
E008	Flash Mob	173	0
E009	Vehicle Unstuck	131	0
E010	Grooming animal	136	0
E011	Make a Sandwich	111	0
E012	Parade	134	0
E013	Parkour	108	0
E014	Repair Appliance	123	0
E015	Sewing	116	0
Rest	Random Other	N/A	3755

a board trick", "feeding an animal", "landing a fish", "wedding ceremony", and "working on a woodworking project"; the remainder of the videos in the test set are random videos not belonging to any of the event categories. The number of videos in each category for train and test is available in Table 1. The contents of these videos are highly variant, for example, the concept "attempting a board trick" includes people skateboarding, snowboarding and surfing, while the "wedding ceremony" varies from a traditional catholic mass, to a Hindi ceremony, to home-made music videos. The analysis presented in this paper is limited to the event kits which are in the training set of the TRECVID MED 2011 data set. Annotators' analyses of the testing set have revealed a huge number of different sound categories some of which a event specific like different tool sounds in woodworking or engine sounds as well a music and different kinds of speech in the videos.

IV. METHODOLOGY

To explore the applicability of speaker diarization to audio concept detection, we applied the ICSI speaker diarization system to the TRECVID MED 2011 data set and analyzed the results produced by speaker segmentation. The basic idea behind this approach is that speaker diarization clusters those segments of an audio stream that exhibit similar acoustic properties into a speaker model. When preprocessing filters such as speech-nonspeech detection are removed from the system, a speaker model does not necessarily represent a speaker, but a low level audio concept. Our motivation was that event specific sounds like a power drill in a video about woodworking, an engine sound in a tire change scenario or clapping sound in a wedding video could be found. Speaker models (which in our case are used as low level audio concepts) are represented by the diarization system as Gaussian Mixture models. A Gaussian Mixture Model (GMM) is a number of Gaussian distributions that describe each feature in the speaker model, as shown in equation (1).

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i N(\vec{x}|\mu_i, \Sigma_i) \tag{1}$$

where $\vec{x}$ is a D-dimension random vector, $N(\vec{x}|\mu_i, \Sigma_i)$, $i = 1, \ldots, M$, are the component densities and w_i, $i = 1, \ldots, M$, are the mixture weights. Each component density is a D-variate Gaussian function of the form with mean vector μ_i and covariance matrix Σ_i (we use diagonal covariance matrix here). The mixture weights are constrained by $\sum_{i=1}^{M} w_i = 1$.

A single feature is represented by a number of Gaussians that are weighted according as to how they influence the overall model. The Gaussian distributions themselves are represented by their mean value and their variance.

In order to match low level audio concepts across training videos and to also be able to classify low level feature models found in testing videos, we have simplified the Gaussian mixture model per speaker to a single vector that consists of the sums of the weighted means and the sums of the weighted variances of each Gaussian. In the remainder of this paper, we will call this vector a simplified supervector, as shown in equation (2)

$$\phi(x) = [\sum_{i=1}^{M} w_i \mu_i \ ; \sum_{i=1}^{M} w_i \Sigma_i] \tag{2}$$

We then clustered the simplified supervectors from all low level acoustic concepts from all video files with a Kmeans approach. The resulting clusters represent abstractions of the simplified supervector for all acoustic low level concepts and can be mapped back to the acoustic low level concepts (speaker models) in each video file by calculating the distance between the abstract simplified supervectors and the individual video's speaker models. By re-mapping the abstract simplified supervectors to the individual speaker models, one can count the overall occurrences of an abstract acoustic low level concept in all videos. We have also counted the number of occurrences of each abstract acoustic low level feature in all videos belonging to each event set. These numbers allow us to compute the normalized frequency of the occurrence of a specific acoustic low level concept sound per event as shown in equation (3).

$$\text{EEH}(c_i, E) = \frac{\sum_k \sum_j n_j P(c_i = c_j | c_j \in D_k \cap D_k \in E)}{\sum_k \sum_j n_j P(c_i = c_j | c_j \in D_k)} \tag{3}$$

where EEH represents expected event histogram and n_j is the occurrence number of c_j in audio clip D_k. $P(c_i = c_j | c_j \in D_k, D_k \in E)$ is the probability of audio term c_i equal c_j given c_j is in the audio clip D_k in the event E,

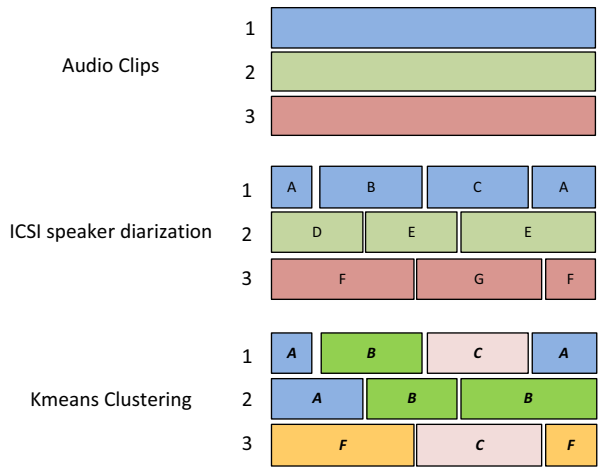


Figure 2. workflow for speaker diarization and clustering

and $P(c_i = c_j | c_j \in D_k)$ is the probability of audio term c_i equal c_j given c_j is from audio clip D_k.

The higher the normalized frequency of a sound belonging to an abstract acoustic low level concept *a* in an event *b* is, the higher is the predictive power of these acoustic low level concepts. To give an extreme example: if a power drill sound could be correctly identified by the system, it would be likely that all occurrences of that sound appear in videos belonging to the category "woodworking", resulting in a normalized frequency of 100%. In other words, whenever a power drill sound occurs, we have a woodworking video. In order to analyze the applicability of speaker diarization, we have extracted those low level acoustic concepts, that have the most predictive per event and examined these closer as discussed in the results section.

The whole workflow is illustrated in figure 2 and consists of two parts: the ICSI speaker diarization system and Kmeans clustering. As shown in the top region of the figure, suppose we have three audio clips: 1, 2, and 3. By applying the ICSI speaker diarization system, each audio clip is segmented into separate chunks. If the chunks are assumed by the system to exhibit similar accoustic properties, then the chunks will be considered to belong to the same speaker. For example, in audio clip 1, there are speakers A, B, and C. Note that each speaker in each audio clip is described by a Gaussian Mixture Model.

Finally, given different speakers from the ICSI speaker diarization system, we use the simplified supervector obtained from weighted mean and variance to represent a speaker. Then, we apply Kmeans clustering to cluster similar speakers among all audio clips. For example, speaker A in clip 1 and speaker D in clip 2 (see figure 2) are clustered together as cluster *A*.

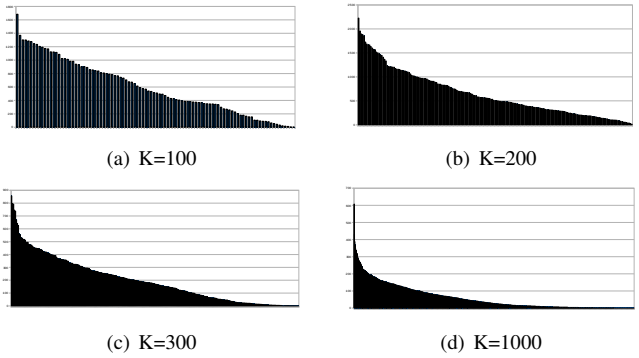
V. Results from Data analysis

The analysis of the distribution of the acoustic low level concepts has revealed a number of acoustic low level concepts that have high predictive power for the abstract concepts (wedding, woodworking, etc.) given in the training set. For five abstract concepts, we found acoustic concepts with a normalized frequency of 50% or greater compared to a chance rate of 1/15 which is 6.7%: 100% for E001, 71.1 for E005, 71.4 for E007, 64.28 for E011, 50 % for E004. I.e. one of these events can be predicted correctly with a probability of 80 % or greater just by the presence of an instance of the respective acoustic low level concept. The event that performed worse in this comparison is Changing a tire, which can only be predicted with a 34 % accuracy based on its dominant low level sound concept. These numbers show that the low level sound concepts generated by speaker diarization are a good discriminator for the higher level concepts of the 15 events given in the NIST TRECVID 2011 data set. Table II shows the normalized frequencies of the top five most predictive acoustic low level concepts for all events in the event training kit. The normalized frequencies are based on kMeans clustering with k=200. The average normalized frequency for the top sound concept for all events was 46.6%. Clustering with k=100 produced an average normalized frequencies of 39.8% for the top sound concept. Clustering with k=1000 produced an average normalized frequency in the 1 % range, likely due to overfitting to specific low level sound concepts from individual video clips.

While some sounds are a very good indicator of a specific event in the training set, their overall frequency of occurrence also has to be considered when using a specific sound concept for event detection. Sound 153 for instance occurs only once in the whole training set. It is a fast gurgling water sound from a video about surfing. Sound 159, the top sound concept in E005 occurs in 21 videos, 12 of which are in E005. Also, multiple speaker models from these videos are matched to that sound concept, which is why the predictive value is 71.4%. We have manually inspected some of the instances of this sound in E005 and all videos where this sound was found in videos from the other event categories. Most instances of this sound concept in E005 are engine sounds of a moderate volume. The sound is also found in E001, E009, E001, E014, and E015. In one instance in E001 it is a harp sound in spherical music, in the other one a compressor sound which is acoustically similar to a moderate engine sound. In E009 it occurs in two videos, in both as an engine sound of racing cars in a distance. In E014 it occurs in only one video, where it is similar to a moderate volume sound of a vacuum cleaner in the background. In E015 it occurs in 4 videos. In 3 of these, it is the sound of a sewing machine and it is acoustically similar to some of the sounds of E005. In the other video

Table II
TOP FIVE SOUND CONCEPTS ACCORDING TO NORMALIZED FREQUENCY

E001	100% (ID 153)	25.8% (ID 117)	25% (ID 130)	19% (ID 189)	18.9% (ID 10)
E002	16.9% (ID 85)	14.8% (ID 108)	14.7% (ID 129)	14.6% (ID 90)	14.2% (ID 184)
E003	40.00% (ID 188)	31.25% (ID 13)	25.58% (ID 18)	22.32% (ID 58)	16.26% (ID 133)
E004	50.00% (ID 47)	41.66% (ID 48)	35.86% (ID 175)	33.33% (ID 94)	30.95% (ID 149)
E005	71.7% (ID 161)	58.3% (ID 88)	25% (ID 130)	24.7% (ID 66)	19.9% (ID 54)
E006	40.0% (ID 188)	14.2% (ID 52)	13.6% (ID 103)	13.3% (ID 4)	13% (ID 71)
E007	71.4% (ID 62)	43.5% (ID 51)	42.8% (ID 199)	41.9% (ID 139)	41.6% (ID 186)
E008	37.4% (ID 178)	29.8% (ID 110)	28.1% (ID 66)	25.0% (ID 130)	21.9% (ID 79)
E009	37.41% (ID 178)	29.81% (ID 110)	28.08% (ID 66)	25.00% (ID 130)	21.91% (ID 79)
E010	25.00% (ID 13)	21.42% (ID 169)	15.00% (ID 157)	14.91% (ID 65)	14.77% (ID 82)
E011	64.28% (ID 169)	36.36% (ID 103)	33.33% (ID 94)	32.98% (ID 70)	23.94% (ID 40)
E012	40.00% (ID 22)	40.00% (ID 156)	34.33% (ID 186)	33.33% (ID 135)	30.00% (ID 176)
E013	23.68% (ID 76)	15.11% (ID 43)	14.24% (ID 2)	13.07% (ID 20)	12.67% (ID 17)
E014	31.42% (ID 177)	30.43% (ID 84)	27.84% (ID 106)	27.08% (ID 98)	24.61% (ID 152)
E015	41.66% (ID 48)	33.33% (ID 94)	19.71% (ID 40)	14.77% (ID 37)	14.28% (ID 62)



(a) K=100 (b) K=200

(c) K=300 (d) K=1000

Figure 3. Distribution Frequency for Kmeans K=100, 200, 300, 1000.

from E015 where it occurs, it is a high frequency sound of rotating equipment. Another observation we made is that there are a number of sound concepts for each specific event class that are not present in any video from that event class. These numbers are E001: 35, E002: 29, E003: 27, E004: 7, E005:20, E006: 11, E007: 32, E008: 11, E009: 28, E010: 14, E011: 15, E012: 27, E013: 23, E014: 66 and E015: 22. Both observations indicate that the presence or absence as well as the number of occurrences of certain low level sound concepts in a video is a predictor for the concept class of that video. Figure 3 show the distribution of low level sound concepts in the test set according to their frequency of occurrence for clustering with k=100, k=200, k=300 and k=1000. The figures suggest that with higher values for k in the clustering, the distribution comes closer to a Zipfian distribution. Zipfian distributions are sometimes connected to the applicability of TF-IDF measures, even though this is controversial from a theoretical perspective [10].

In addition to the analysis of the top performing sound concepts, we annotated and studied the distributions of seven sound sets, without prejudice towards how well they performed. In sound 5 we have a fairly poor discriminator, although by no means useless: it occurs in all of the event categories, with 41% of it's occurrences being in E004, E005 and E008. Across the various event classes, it was observed to contain fairly similar sounds: that of guitar music and singing being most common, and where it is most highly discriminating. The cluster also contains machine noises, speech, and the sound of bacon cooking. Sound 8 on the other hand only has one event with a greater than 10% chance of occurrence, E014 at 16% and seems to be a cluster of sounds that include music with percussive elements. In E014 this translated to music with tools being used, but in other places it was bass heavy techno, or cars moving at high speed. With sound 76 we had a sound which to a human annotator seems relatively unremarkable, being mostly instrumental music; however this sound was only detected in 25 of the videos, and not detected at all in six of the event classes.

VI. CONCLUSION

This more in depth analysis of how our system has created the sound clusters is illuminating: the machine learning system is creating categories that while clearly sensical for it, and useful in classifying the events, are clearly not what a human annotator would create. This forces us to reevaluate the utility of our standard annotation approaches while preparing the data for this sort of system, since what the system is discovering, and finding useful is quite different from what a human annotator might create; it seems unlikely that a human annotator would put guitar music in the same category as bacon cooking. This dichotomy will hopefully prove to be advantageous in the future, if we can develop a system that can combine the human understanding of sound meanings with the automatic segmentations which we have been using.

The distribution analysis of the clusters generated by the two processing steps speaker diarization and clustering has shown that the distribution of low level sound concepts found by speaker diarization differs between videos belonging to different classes. It is hence safe to assume that low

level sound concepts can be used for video classification. One advantage of using a representation of videos at the level of the distribution of low level sound concepts is that low level sound concepts deliver an abstract representation. Classification at this level does hence require only a small amount of data. One video can be described by a vector with 200 or 300 positions instead of a framewise representation of raw features. Preliminary machine learning experiments with this high level representation have confirmed our estimates. In the future we will design a video concept classification system based on the low level sound concepts found by speaker diarization.

VII. Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmenta l purposes notwithstanding any copyright annotation thereon. The views and conclusion contained herein are those ot the authors and should not be interpresented as necessarily representing the official policies or endorsement, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Governement.

References

[1] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, no. 1, pp. 1–19, 2006.

[2] Cees G. M. Snoek and Marcel Worring, "Concept-based video retrieval," *Fundamental Trends in Information Retrieval*, vol. 2, no. 4, pp. 215–322, 2009.

[3] HD Wactlar, T. Kanade, MA Smith, and SM Stevens, "Intelligent access to digital video: Informedia project," *Computer*, vol. 29, no. 5, pp. 46–52, 1996.

[4] Huan Li, Lei Bao, Zan Gao, Arnold Overwijk, Wei Liu, Long fei Zhang, Shoou-I Yu, Ming yu Chen, Florian Metze, and Alexander Hauptmann, "Informedia@trecvid 2010," in *Notebook for NIST's TREC Video Retrieval Evaluation 2010*, 2010.

[5] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang, "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching," in *NIST TRECVID Workshop*, 2010.

[6] David Imseng and Gerald Friedland, "Robust speaker diarization for short speech recordings," in *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, 12 2009, pp. 432–437.

[7] Chuck Wooters and Marijn Huijbregts, "Multimodal technologies for perception of humans," chapter The ICSI RT07s Speaker Diarization System, pp. 509–519. Springer-Verlag, Berlin, Heidelberg, 2008.

[8] G. Friedland and O. Vinyals, "Live speaker identification in conversation," in *Proceedings of the ACM International Conference on Multimedia*, October 2008, pp. 1017–1018.

[9] Yan Huang, Oriol Vinyals, Gerald Friedl, Christian Mller, Nikki Mirghafori, and Chuck Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," in *ASRU*, 2007.

[10] Stephen Robertson, "Understanding inverse document frequency: On theoretical arguments for idf," *Journal of Documentation*, vol. 60, pp. 2004, 2004.