EXTRACTING SPOKEN AND ACOUSTIC CONCEPTS FOR MULTIMEDIA EVENT DETECTION

Julien van Hout[1], Murat Akbacak[2], Diego Castan[3], Eric Yeh[1], Michelle Sanchez[1]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, U.S.A.
[2] Microsoft, Sunnyvale, CA, U.S.A.
[3]ViVoLab - I3A, University of Zaragoza, Zaragoza, Spain

ABSTRACT

Because of the popularity of online videos, there has been much interest in recent years in audio processing for the improvement of online video search. In this paper, we explore using acoustic concepts and spoken concepts extracted via audio segmentation/recognition and speech recognition respectively for Multimedia Event Detection (MED). To extract spoken concepts, a segmenter trained on annotated data from user videos segments the audio into three classes: speech, music, and other sounds. The speech segments are passed to an Automatic Speech Recognition (ASR) engine, and words from the 1-best ASR output, as well as posterior-weighted word counts collected from ASR lattices, are used as features to an SVM based classifier. Acoustic concepts are extracted using the 3-gram lattice counts of two Acoustic Concept Recognition (ACR) systems trained on 7 and 22 classes. MED results are reported on a subset of the NIST 2011 TRECVID data. We find that spoken concepts using lattices yield a 15% relative improvement in Average Pmiss (APM) over 1-best based features. Acoustic concepts with 22 classes gave a 30% relative gain in APM over using 7 classes. Lastly, while our best acoustic and spoken concepts yield similar performance individually, we obtain a 28% relative APM improvement after score-level fusion of both concept types.

Index Terms— Multimedia event detection, segmentation, speech recognition, acoustic event recognition, lattice N-gram counts

1. INTRODUCTION

Due to the popularity of online videos, there has recently been significant interest in multimedia analysis. Features in the video imagery play a significant role in determining the content. However, as multimedia event detectors go beyond retrieving simple events (e.g., detecting a baseball game), and move towards specific and hard-to-detect events, such as “home run in a baseball game”, audio and spoken content features become more important as they provide supplemental information to image/video features. In the above example, analysis of the frame-level imagery may determine that the setting is a baseball game, but without the capability to capture cheering in the audio or spoken keywords, it would be significantly more difficult to discriminate between an uneventful game and one with a home run.

In the last two decades, speech recognition has been applied to constrained domains (e.g., broadcast news, telephone conversations, meetings) where the test collection was fairly homogeneous in terms of acoustic conditions and lexical content. Recently, with the expanding size of user-submitted online videos, ASR researchers started evaluating state-of-the-art ASR systems on these heterogeneous data collections. In addition to moving towards more heterogeneous collections, ASR started becoming an important part of multimedia projects since spoken content provides information that is both discriminative and complementary to video imagery features. In the past, ASR content has been used for applications such as spoken document retrieval [1] or topic classification [2]; however, only a few studies [3, 4] reported using ASR for MED on heterogeneous collections such as user-submitted videos. Our main contribution over [3, 4] lies in using ASR lattices to compute expected word-counts that provide more robust features than counts extracted form the N-best. We also try system combination of spoken and acoustic concepts to leverage information from non-spoken content.

The MED track under the NIST Text Retrieval Conference on Video Information Retrieval (TRECVID) evaluations attempts to build technology that enables search for specific events in user-submitted quality videos [5]. To handle this challenging problem, systems typically extract a set of heterogeneous low-level audio, visual, and motion features, as well as higher-level semantic content in the form of audio and visual concepts, spoken text, and video text. Each type of feature or semantic content is extracted by one or more event classifiers that are trained with examples from the training set. As we mentioned above, spoken concepts based on ASR could play a key role in retrieving complex events. Yet, several challenges arise when ASR systems are employed in such domains. First, detecting speech becomes harder as sometimes speech is overlapped with background noises or music. Second, due to the variety in acoustic conditions (e.g., different recording conditions and equipment, varying quality of speech, background noise overlapped with speech), a mismatch occurs between training and test conditions, affecting ASR accuracy. A similar mismatch is expected for the language modeling component of ASR as well. As the data collections are getting multilingual, a mismatch in language itself is unavoidable. Similarly to what has been done in spoken document retrieval or topic classification, ASR N-best hypotheses have been used as features for MED classifiers in replacement to using only the 1-best ASR output [3, 4], along with some keyword expansion techniques [6]. In addition to word hypotheses, some studies (e.g., [7]) explore using multilingual phone recognizers to model spoken content, especially because the spoken content in some MED datasets contains speech from different languages.

In this study, we first present both the segmentation and ASR improvements obtained by training a 4-class speech segmentation on LDC annotated TRECVID data. We then present the ASR and Acoustic Concept Recognition (ACR) [8] systems used to extract spoken and acoustic concepts. We propose extracting expected counts from the ASR and ACR lattices as a more robust measure

Murat Akbacak and Diego Castan were at SRI International when this work was done.

of word appearance than 1-best counts, and use them as features for MED. To model spoken concepts, we consider both a maximum entropy classifier and a linear SVM. Various feature processing techniques are presented: normalization at the video level, word counts weighting, and feature dimension reduction by word selection and stemming. We present MED results on a subset of the TRECVID 2011 MED training data and compare the performance of the proposed spoken concepts with broad and specific acoustic concepts, and with the system obtained by applying score-level fusion of both concept types. An overview of our system is shown in Figure 1.

The paper is organized as follows. In Section 2, we describe the updates made to the segmentation module. The ASR system is presented in Section 3. In Section 4 we then discuss the lattice-based feature extraction for spoken and acoustic concepts, and in Section 5 we discuss the experimental setup and results. This is followed by conclusions and future work in Section 6.

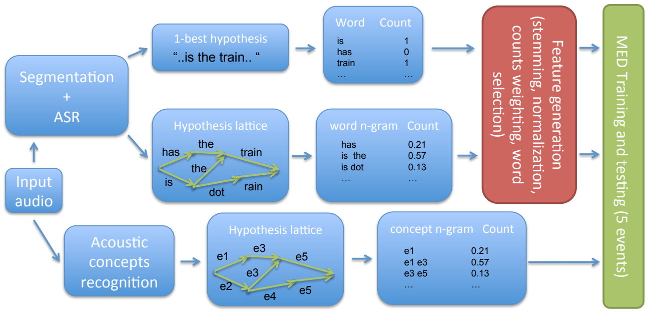

Fig. 1: Extracting acoustic and spoken concepts as features for MED

2. SEGMENTATION

Since audio from multimedia events is so heterogeneous in nature, a good segmenter is essential in order to determine which segments need to be fed to the ASR system. Here, we briefly describe how the segmentation system was designed, trained and evaluated on audio extracted from the videos.

We build a segmenter with four classes: speech, music, noise, and pause. Each class is modeled by a 3-state HMM-GMM with self loops and 256 fully tied gaussians. The models were trained using Mel-Frequency Cepstral Coefficients (MFCC) features. The training set includes 619 videos internally annotated as pure music, 2658 segments internally annotated with noise, and 544 videos annotated with speech from part 1 and 2 of LDC's release of pilot transcripts for TRECVID MED-11 video files (LDC2012E08). The pause model was trained using audio and transcripts from the Wall-Street Journal corpus. We build the decoding network as a loop of the four trained HMMs, where we concatenated 2 HMMs for each of the first three labels (speech, music, noise) and 4 HMMs for the pause label, thus enforcing a minimum duration constraint. The insertion penalties are set to 0, -10^3, -10^5, -10^5 for pause, speech, music and noise, respectively. Finally, segments of speech exceeding 50 seconds in duration are split into smaller pieces to limit the ASR runtime.

A validation set was designed using the remaining 627 annotated videos from part 3 of LDC2012E08, totaling about 19 hours of audio of which 6 hours are speech. A confusion matrix is shown in Table 1. The overall speech false alarm rate is 25.3%, while the false reject rate is 22.3%. It is worth noting that noise is in many instances recognized as speech while the opposite is less frequent. This is likely due to the many overlaps between speech and noises in the training data, and is a major problem with TRECVID MED data.

We also evaluated the segmentation performance jointly with ASR in order to assess how the segmentation performance impacts the Word-Error Rate (WER). Speech segments from the 627 abovementionned videos are processed through the ASR engine described in Section 3. The WER is shown in Table 2 where three different segmenters are compared. The baseline segmenter is a classic GMM-based Speech Activity Detection (SAD) system trained on meetings data. The *Oracle* segmenter uses the ground truth segmentation from LDC transcripts and provides a perfect segmentation. We observe that the proposed system performs significantly better than the baseline system, highlighting the importance of training the segmentation directly on the multimedia audio. Interestingly, because of the difficulties of ASR on such data, the best possible segmentation provides only a 2% Precision improvement over the proposed system.

Table 1: Confusion matrix of segmentation on 627 videos

Reference	Hypothesis				# frames
	Speech	Music	Noise	Pause	
Speech	77.7%	12.4%	8.9%	0.9%	2,092,691
Music	12.5%	63.5%	19.6%	4.4%	3,030,826
Noise	37.1%	18.7%	43.3%	0.9%	1,377,194
Pause	64.1%	11.6%	23.4%	0.9%	582,597

Table 2: Impact of segmentation on ASR performance.

Segmentation	%Corr	%Err	%Sub	%Del	%Ins
Baseline	24	105	32	44	29
Proposed	28	83	32	40	11
Oracle	30	78	35	36	7

3. AUTOMATIC SPEECH RECOGNITION

Here, we describe the ASR system that was chosen to generate spoken concepts. For this task, we ran several in-house English ASR systems trained on data with channel and speaker characteristics relating as much as possible to those of the observed TRECVID MED data. The ICSI-SRI 2006 meetings recognition system [9] adapted to Single Distant Microphone (SDM) data yielded the best performance in terms of WER on a test set of TRECVID MED audio and thus was picked to generate the proposed spoken concepts. In order to limit run-time, we only used 1-best and lattices from first-pass decoding using a within-word triphone MFCC model. The MFCC models used 12 cepstral coefficients, energy, first-, second-, and third-order differences features, and 2×5 voicing features over a 5-frame window. This system uses both speaker clustering, Vocal-Tract Length Normalization (VTLN), and Heteroscedastic Linear Discriminant Analysis (HLDA). After HLDA, a 25-dimensional Tandem/HATs feature vector estimated by multilayer perceptrons (MLPs) was appended. The MFCC recognition models were derived from gender-dependent Conversational Telephone Speech (CTS) models in our RT-04F system, which had been trained with the minimum phone error (MPE) criterion on about 1400 hours of CTS data. These models were then adapted to 100 hours of SDM data using a standard maximum likelihood (ML) Maximum A Posteriori (MAP) procedure. During decoding, the MFCC models were adapted through maximum-likelihood linear regression (MLLR) using a phone-loop model as reference. Lattices were then generated using a multiword bigram language model (LM) linearly interpolated from component LMs trained on various corpora. More details on the ASR system can be found in [9]

4. EXTRACTING ACOUSTIC AND SPOKEN CONCEPTS

Here, we present our approach to using acoustic and spoken concepts as features for the MED task. First, we present a summary of the ACR system from [8], that was used to recognize either 7 or 22 acoustic concepts through the videos. From the ACR lattices, we extract 3-gram expected counts and use them as features for MED. Then, we present our approach to generating spoken concepts from the ASR output by using both the 1-best output and the ASR lattices. In this part, we also present the feature processing techniques that we considered to normalize the features and reduce the feature dimension through stemming and word selection.

4.1. Extracting acoustic concepts

Given that one of the long-term goals of TRECVID MED track is to detect multimedia events based on high-level features, we created a set of acoustic categories that we call *acoustic concepts*. These categories were chosen to be useful in discriminating the video event classes while being clear to both audio annotators and potential users of the MED system. These acoustic concepts are divided into five broad classes, which are in turn split into 20 specific classes as shown in Table 3. Two classes for speech and music are added to both of these lists, resulting in 7 broad classes and 22 specific classes. Additional details about designing and annotating the acoustic concepts can be found in [10]. We perform supervised training using both the 7 and 22 classes. While the ACR system [8] is not the main focus of this paper, it is summarized here for the sake of completeness. The ACR based MED system will be used in our experiments as a comparison point and as a second system with which the proposed MED system based on spoken concepts will be combined.

We extract 16 MFCC (including C0), computed in 25ms frame size with a 10ms frame step, their Δ and $\Delta\Delta$ and apply cepstral mean subtraction on each video as a normalization step. We then compute the mean and standard deviation over 1 second windows with an overlap of 0.75 second and use those 96-dimensionnal features to represent the acoustic concepts. Each concept is then modeled using a one-state HMM-GMM with 256 Gaussians. Using the HTK HMM toolkit [11], we build a grammar-free decoding network that is used to produce lattices encoding multiple recognition hypotheses and with their acoustic likelihoods.

The motivation to using ACR for MED is that sequences of acoustic concepts are believed to be strongly indicative of a specific multimedia event. Similar approaches have been applied successfully in the past to determine different languages [12, 13] or to identify dialects [14]. The ACR lattices are therefore used to compute the expected count of every concept N-gram for N as high as 3. We vectorize the N-grams to generate an SVM feature vector. Expected counts of N-grams can be easily understood as an extension of standard counts. Given a hypothesized string of concepts, $C = c_1, \cdot\cdot\cdot, c_n$, the count for a given N-gram d_i is $count(d_i|C)$, the number of occurrences of d_i in the sequence C. To extend counts to a lattice L, we find the expected count over all possible concept sequences:

$$count(d_i|L) = E_C[count(d_i|C)] = \textstyle\sum_{C\epsilon L} p(C|L)count(d_i|C)$$

4.2. Extracting spoken concepts

Spoken concepts are high-level features characterizing the spoken content of a video, which can provide highly valuable information for MED. This section describes the various approaches that we considered to extract feature vectors from the ASR output.

First, we compared using the 1-best ASR hypothesis or lattice-based expected word counts. The 1-best ASR output is the most likely word sequence extracted from the lattice, while expected word counts are computed similarly to the concepts N-gram counts with $N = 1$. Because of the relatively low accuracy (28%) of the 1-best ASR, lattice-based counts are expected to be more reliable than 1-best counts. For each video, these counts are aggregated into a feature vector of dimension 54484, the size of the ASR vocabulary.

Table 3: Broad and Specific acoustic concepts

5 Broad Concepts	20 Specific Concepts	
	Air traffic	Individual applause
	Birds	Individual yells
Crowd/audience	Crowd applause	Large crowd
Animal sounds	Crowd cheers	Scraping/Sanding
Repetitive sounds	Crowd laughter	Sewing
Machine noise	Crowd yells	Skateboard
Environmental sound	Farm animals	Small party
	Ground traffic	Water running
	Hammer	Water splashing
	Wind	Home appliances

Second, we looked at normalizing the counts of each videos by the number of spoken words. This normalization ($\mathbf{norm}$) allows the counts to be independent of the video length and amount of speech. While such a property is desirable for heterogeneous data, un-normalized counts intrinsically capture the duration of speech, which can also be a discriminative feature for some MED events.

Third, we tried various word counts weighting techniques. Since some words are inherently more frequent than others, their counts can be several orders of magnitude larger than counts of rarer but potentially discriminative words. We tried several weighting schemes by which we boost counts of infrequent words over frequent ones. The first weighting approach ($\mathbf{W_{Log}}$) maps raw expected counts $c(w|L)$ to log counts $c_{log}(w|L)$ as follows: $c_{log}(w|L) = log(c(w|L) + f)$ where f is a flooring parameter that was optimized to 10^{-4} and helps limit the impact of infrequent words. The second approach ($\mathbf{W_{tf}}$) we try is to weight each word count by the inverse average count for this word over the corpus, at some power α:

$$c_{tf,\alpha}(w|L) = \frac{c(w|L)}{(\sum_{M\in Docs} c(w|M))^{\alpha}}$$

A third approach ($\mathbf{W_{Logtf}}$) combines the two previous schemes as follows: $c_{logtf,\alpha}(w|L) = log(c_{tf,\alpha}(w|L) + f)$.

Fourth, we looked at techniques to reduce the feature dimension. A first approach, stemming, is a common technique applied in information retrieval that we used to reduce inflected or derived words to their stem (base or root form, but not necessarily morphological root). We used the Porter stemmer [15] to reduce the vocabulary size to 34489 words, and summed the counts of words mapping to the same stem. As a second approach to dimension reduction, we looked at using only a subset of discriminative words through word selection. The following data-driven criteria were considered:

A large count over all videos

B large mean count for at least one event

C low within-event entropy from the 15k top ranking words in **B**

D 500 top-scoring words from a logistic regression classifier

E 500 words extracted from the MED event descriptions and semantic expansion on those using web-based search.

List **E** combines the top TF-IDF ranked tokens from the TRECVID event descriptions with scored tokens derived using a semantic similarity technique, Explicit Semantic Analysis (ESA) [16]. ESA was used to identify the top 100 most similar Wikipedia articles, and the tokens in the article titles were collected and scored using TF-IDF. In Table 4, results on these lower dimensional features are reported along with the lowest dimension that provided little loss of accuracy, whenever possible.

5. EXPERIMENTAL RESULTS

MED experiments are run on a subset of the NIST 2011 TRECVID training data comprising 6016 videos for training and 1842 videos for testing. This subset was chosen so that all of the videos had some hypothesized speech. The following five multimedia events are detected: Attempting a board trick (*E001*), Feeding an animal (*E002*), Landing a fish (*E003*), Wedding ceremony (*E004*) and Working on a woodworking project (*E005*). The number of these varies from 53 to 82 for training and from 18 to 28 for testing across the events. For our MED experiments, we used either a logistic regression-based classifier with L^2 regularization (MaxEnt) or a linear SVM using the SVM-light implementation from [17]. MED results are shown in Table 4 for each of the five events in terms of Average P_{miss} (APM), which measures the area under a Detection-Error Tradeoff curve.

For spoken concepts, the two runs with a MaxEnt classifier confirm that using lattice-counts performs better than using 1-best counts (0.47 to 0.39 APM). Further, we see that using a linear SVM provides substantial gains over using a MaxEnt classifier (from 0.39 to 0.34 APM). An SVM is therefore used for all remaining experiments. We observe that stemming the words and adding the counts to those sharing the same stem brings a rather small improvement since the APM goes down by 0.1 when stemming is applied, with and without log-weighting. We observe a large increase in APM when applying word-count normalization by the number of spoken words, both without weighting (0.21 to 0.46), and in combination with the three proposed weighting schemes. This seems to validate our hypothesis that the number of spoken words is an essential feature in discriminating videos belonging to MED events. We also see that log weighting gives better results (0.21 APM) than TF weighting with $\alpha = 1$ (0.44) or $\alpha = 0.5$ (0.37), or a combination of the two (0.22 and 0.23). None of the word-selection criteria performed better than the 34k vocabulary baseline (0.21 APM), but criteria A and B performed almost as well with a much reduced vocabulary (0.22). By comparison, criteria C, D and E, which ignore high-frequency terms to focus on a priori discriminative words, perform poorly (0.30 to 0.36). This shows that using at least 1000 high-frequency words is essential to provide good accuracy, but supplemental discriminative words from lists C, D and E could still be used in combination.

The MED system based on ACR performed significantly better with 22 events (0.21 APM) than with 7 events (0.30), reaching a level of performance similar to that of the best ASR-based MED system.

In order to leverage information from both Acoustic and Spoken concepts, we perform score-level fusion of these two or three systems by normalizing their prediction scores to have zero-mean and unit variance and adding them with equal weighting. The best performing combination was obtained by combining the ASR MED system with the 22-concepts ACR MED system. The combined system performed better than both of the original systems for all five events, and provided a relative 28% improvement in APM (from 0.21 to 0.15) in averaged over all events. This result shows that acoustic and spoken concepts capture a different kind of information that can be easily combined to build a significantly more robust MED system.

Table 4: Average-P_{miss} by event for the proposed MED systems

System	Event					
MaxEnt classifier	E001	E002	E003	E004	E005	Avg.
1-best no-stem	0.30	0.35	0.48	0.62	0.58	**0.47**
Lattice no-stem	0.31	0.36	0.47	0.33	0.46	**0.39**
ASR Lattice SVM	E001	E002	E003	E004	E005	Avg.
no-stem	0.30	0.37	0.50	0.23	0.29	**0.34**
stem	0.30	0.38	0.45	0.23	0.28	**0.33**
no-stem W_{Log}	0.26	0.27	0.23	0.20	0.16	**0.22**
stem W_{Log} (**1**)	0.26	0.27	0.21	0.19	0.14	**0.21**
stem $W_{tf,0.5}$	0.31	0.41	0.43	0.30	0.38	**0.37**
stem $W_{tf,1}$	0.38	0.35	0.49	0.44	0.54	**0.44**
stem $W_{logtf,0.5}$	0.26	0.24	0.26	0.17	0.17	**0.22**
stem $W_{logtf,1}$	0.28	0.21	0.28	0.17	0.18	**0.23**
stem norm	0.54	0.44	0.56	0.45	0.33	**0.46**
stem W_{Log} norm	0.57	0.43	0.48	0.58	0.51	**0.51**
stem $W_{tf,0.5}$ norm	0.51	0.46	0.45	0.50	0.48	**0.48**
stem $W_{tf,1}$ norm	0.55	0.41	0.55	0.54	0.50	**0.51**
stem W_{Log} A-1000	0.26	0.28	0.21	0.18	0.15	**0.22**
stem W_{Log} B-5000	0.27	0.28	0.21	0.19	0.13	**0.22**
stem W_{Log} C-5000	0.34	0.33	0.37	0.30	0.40	**0.35**
stem W_{Log} D-500	0.33	0.31	0.30	0.27	0.31	**0.30**
stem W_{Log} E-500	0.30	0.35	0.43	0.38	0.33	**0.36**
ACR Lattice SVM	E001	E002	E003	E004	E005	Avg.
7 concepts (**2**)	0.22	0.41	0.32	0.25	0.28	**0.30**
22 concepts (**3**)	0.17	0.30	0.23	0.15	0.20	**0.21**
Fusion ASR + ACR	E001	E002	E003	E004	E005	Avg.
(**1**) + (**2**)	0.21	0.23	0.22	0.17	0.15	**0.20**
(**1**) + (**3**)	0.16	0.23	0.15	0.11	0.13	**0.15**
(**2**) + (**3**)	0.15	0.32	0.22	0.16	0.18	**0.20**
(**1**) + (**2**) + (**3**)	0.14	0.23	0.16	0.12	0.13	**0.15**

6. CONCLUSIONS AND FUTURE WORK

We present a robust approach to building an audio-based Multimedia Event Detection system using spoken and acoustic concepts. Acoustic concepts are generated by performing ACR [8] on 7 or 22 general classes and converting the recognition lattices into a vector of expected 3-gram counts. A similar approach is used to extract spoken concepts. We first apply speech/non-speech segmentation and feed speech segments to an ASR engine that produces lattices from which we computed expected word counts. We find that stemming the lattices and using a vector of log-counts to train a linear SVM gives the best results. On a subset of the TRECVID MED 11 data, we find that spoken and acoustic concepts both perform similarly on average for an APM of 0.21 across the five events. After score-level system combination of both MED systems, we leverage the complementarity of both approaches and obtain a significant 28% relative decrease in APM. Future work on spoken concepts will look at improving the ASR by doing acoustic and language adaptation using TRECVID audio. We plan to improve our ASR-based features through better word-selection (mixing lists of frequent words like A and B with discriminative ones like C, D and E) and by enriching 1-gram word counts with frequent 2-gram and 3-gram counts. Finally, we will investigate more complex fusion techniques at the score level, by training event-specific fusion weights for each system, and at the feature level by training a single classifier with both concepts types.

7. ACKNOWLEDGMENTS

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon.
Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

8. REFERENCES

[1] M. Akbacak, *"Robust spoken document retrieval in multilingual and noisy acoustic environments"*, Ph.D. thesis, University of Colorado at Boulder, 2009.

[2] D. Hillard, "Topic Classification for Conversational Speech using Support Vector Machines and Latent Semantic Analysis," 2008.

[3] P. Natarajan, P. Natarajan, V. Manohar, S. Wu, S. Tsakalidis, S.N. Vitaladevuni, X. Zhuang, R. Prasad, G. Ye, D. Liu, et al., "BBN VISER TRECVID 2011 multimedia event detection system," in *Proc. of NIST TRECVID Workshop*, 2011.

[4] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, and R. Prasad, "Multimodal feature fusion for robust event detection in web videos," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1298–1305.

[5] D. Scott, J. Guo, C. Foley, F. Hopfgartner, C. Gurrin, and A.F. Smeaton, "TRECVid 2011 Experiments at Dublin City University," *TRECVID 2011*, vol. 12, 2011.

[6] S. Tsakalidis, X. Zhuang, R. Hsiao, S. Wu, P. Natarajan, R. Prasad, and P. Natarajan, "Robust Event Detection From Spoken Content In Consumer Domain Videos," in *Proceedings of Interspeech, Portland, Oregon, USA*, 2012.

[7] Q. Jin, P. F. Schulam, S.h Rawat, S. Burger, D. Ding, and F. Metze, "Event-based Video Retrieval Using Audio," in *Proceedings of Interspeech, Portland, Oregon, USA*, 2012.

[8] D. Castan and M. Akbacak, "Using Acoustic Concept Recognition Lattices for Multimedia Event Detection," in *Submitted to ICASSP 2013*.

[9] A. Janin, A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, J. Frankel, and J. Zheng, "The ICSI-SRI spring 2006 meeting recognition system," *Machine Learning for Multimodal Interaction*, pp. 444–456, 2006.

[10] S. Pancoast, M. Akbacak, and M. Sanchez, "Supervised acoustic concept extraction for multimedia event detection," in *ACM Multimedia Workshop*, 2012.

[11] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," *Cambridge University Engineering Department*, vol. 3, 2002.

[12] FS. Richardson and WM. Campbell, "Language recognition with discriminative keyword selection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[13] WM. Campbell, F. Richardson, and D. Reynolds, "Language Recognition with Word Lattices and Support Vector Machines," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[14] M. Akbacak, D. Vergyri, A. Stolcke, and Scheffer N., "Effective Arabic dialect classification using diverse phonotactic models," in *Interspeech*, 2011.

[15] M.F. Porter et al., "An algorithm for suffix stripping," 1980, pp. 130–137, Program.

[16] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proceedings of the 20th international joint conference on Artifical intelligence*, 2007, pp. 1606–1611.

[17] T. Joachims, "SVMlight: Support Vector Machine," *SVM-Light Support Vector Machine http://svmlight.joachims.org/, University of Dortmund*, vol. 19, pp. 4, 1999.