Audio Self Organized Units for High-level Event Detection

Xiaodan Zhuang, Shuang Wu, Pradeep Natarajan, Rohit Prasad, Prem Natarajan

Speech, Language and Multimedia Business Unit
Raytheon BBN Technologies
{xzhuang, swu, pradeepn, rprasad, pnataraj}@bbn.com

Abstract

High-level multimedia event detection aims to identify videos containing a target event. Recent approaches leveraging audio information for this task fall into two broad categories. The first corresponds to holistic bag-of-words approaches based on frame-level descriptors. These are effective for classification, but hard for humans to interpret. The second corresponds to approaches that build a limited set of mid-level concept detectors trained using large amounts of annotated data. Such approaches do not scale easily for large scale tasks with heterogeneous data. We explore using audio Self Organized Units (SOU) to capture mid-level segmental information in a completely unsupervised fashion, and devise various features based on the SOU decoding process on each video. We train BBN's speech SOU system on unannotated web audio data. A multi-pass adaptive decoder from the BBN speech recognition system is engaged to decode audio data using the HMM-based audio SOUs. We devise various vector representations from the audio SOU lattices and from the constrained maximum likelihood linear regression adaptation matrices at different stages of the decoding. High-level event detection using these representations shows promising results on the benchmark 2011 TRECVID Multimedia Event Detection dataset. Furthermore, the audio SOUs provide potential for human interpretable features.

Index Terms: audio search, lattice, segmentation

1. Introduction

High-level multimedia event detection in the TRECVID Multimedia Event Detection task aims to identify video clips containing a target event [1]. Different from applications in constrained domain, this task calls for generalizable approaches that can handle highly heterogeneous data and diversified event types.

The audio stream of the video provides complementary information. Spoken content can be highly related to the topic. General acoustic environment helps distinguish the context of the video, such as a cheerful indoor party, a quiet classroom and a noisy downtown streetview. Mid-level information based on cues at segmental levels accommodates the heterogeneous nature of the data, focussing on a scale between local frames (e.g., 25ms) and the whole video. In this scale also exist salient temporally local occurences that can be highly informative, such as gunshot noise, screaming and birthday song melody.

Acknowledgement: Supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. 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.

We identify several challenges in engaging audio information for multimedia event detection. First, all these different forms of audio information are often obtained from very different systems [2, 3], e.g., large-vocabulary speech recognition, holistic bag-of-words approach based on frame-level descriptors, and detectors for specific salient events. This significantly complicates the overall system, and sometimes suffers from inaccurate audio segmentation, such as in speech activity detection and localized event detection. Second, the mid-level information can be handled by specific event detectors aiming at salient events in the audio stream [4], such as cheering in sports events [5]. However, a small set of hand-picked events tuned for specific high-level events can hardly support detection of diversified event types, and the annotations required for training are labor intensive. Besides, the audio information useful for high-level event detection is not restricted to such salient events even when they are present. Third, the effective bag-of-words approach based on frame-level, e.g., every 10ms, descriptors [6] provides good performance on identifying the event type of the whole video, but lends very limited human-interpretable insight behind the decision.

In this work, we explore the efficacy of audio Self Organized Units (SOU) for high-level event detection. We train BBN's speech SOU system [7] on unannotated general audio data. A multi-pass adaptive decoder from the BBN Byblos speech recognition system [8] is engaged to decode general audio data using the HMM-based audio SOUs. We devise various vector representations from the audio SOU lattices and from the maximum likelihood linear regression (MLLR) adaptation matrices at different stages of the decoding. This approach has merits that address the challenges of the task. The audio SOUs capture segment-level cues with varying lengths, via the state sequence of the HMMs determined through the multiple-pass decoding and via the sequential information retained in the SOU decoding lattices. The approach is completely unsupervised and can be efficiently applied to different event types without the need to manually define the mid-level concept set and obtain annotation for training detectors. The different representations proposed can capture rich information about the video even with a very small set of audio SOUs. Finally, this approach can potentially provide semantically meaningful summary of the audio, via the salient sequences of audio SOUs.

We perform high-level event detection using various representations based on the audio SOUs. The proposed methods show promising results on the benchmark 2011 TRECVID Multimedia Event Detection dataset. Besides, we demonstrate that the audio SOUs capture audio cues at a temporal scale much larger than frame-level descriptors.

2. Relationship with Previous Work

Since 2010, the BBN HMM-based Self Organized Units approach has been studied for various speech content analysis tasks, such as keyword discovery, spoken topic identification [7] and salient spoken phrase discovery [9]. All these work focus on the analysis of spoken content, and it is not clear whether the HMM-based SOUs can be used for effective general audio analysis. Compared to speech with intrinsic temporal structure, general audio is even more heterogeneous, and our understanding about its temporal structure is more limited. Further, while both use the same decoding procedure, the representation generation based on the decoding output and the classifier setups differ between the speech SOU work [7] and this work.

Some parallel recent work explored a similar approach of using unsupervised HMMs for audio classification [10]. Our work differs from theirs. First, their work focuses on using the 1-best hypothesis transcript, while we consider the audio SOU lattices from the decoding output. Second, once the HMMs are trained, their work performs a one pass Viterbi decoding, while we engage a multi-pass decoder from a state-of-the-art speech recognition system, that enables generation of much richer representations from the decodig process. Third, their work focuses on a closed-set classification problem. They use class-dependent language models to perform multiple parallel decoding, and concatenate class-dependent acoustic scores and language model scores as the representation used in a linear classifier. Our audio SOU approach performs class-independent decoding and extracts representations from different parts of this decoding process. Our approach can support detection of more diversified event types, and enables applications where the target class is not known in representation generation, and only available for training the final classifier.

3. Self Organized Units for General Audio

The BBN SOU models are a set of HMMs trained in a fully unsupervised fashion. In particular, the SOU set is first identified by clustering audio segments delimited by spectral discontinuity. A segmental Gaussian mixture model (GMM) is trained on these clusters and produces the tokenized data for initializing the HMM-based SOUs. Then the HMMs are trained in a state-of-the-art speech recognition system in an iterative fashion that does not require annotated data. In this section, we briefly introduce the training of the SOUs on general audio. All the components are from the BBN Byblos speech recognizer [8] and similar to the SOU training for speech processing [7, 9].

3.1. Identify the SOU set

A set of SOUs are used to model all the audio data, regardless of whether it is speech, music, or other general audio content. Therefore, these SOUs should not only focus on segments that stand out within the video.

Given this consideration, we devise an initial tokenizer to identify the set of SOUs that will be modeled by HMMs. Instead of relying on a tokenizer trained on annotated data, a fully unsupervised approach based on the segmental GMM is adopted, involving the following steps.

- Identify spectral discontinuities in the audio signal.
- Fit each delimited audio segment with a quadratic polynomial trajectory in the cepstral space.
- Calculate pairwise distances between audio segments according to the area between the polynomial trajectories.
- Group the audio segments into clusters, each corresponding to one unique SOU.
- Train a segmental GMM according to the set of audio segments and their corresponding cluster ids.
- Use the segmental GMM to propose the most likely mixture id for each audio segment.

The above labeled audio data will be used for initialization in HMM-based SOU training. The set of SOUs are identified according to the clusters in this procedure.

3.2. Hidden Markov Models for SOUs

The SOU HMMs are trained using the BBN Byblos speech recognition system [8], with each SOU corresponding to a single-phoneme word.

Standard cepstral feature normalization and Heteroscedastic Linear Discriminant Analysis feature transformation are performed as in speech processing for automatic speech recognition (ASR). Context-dependent triphone and quiphone models, various parameter tying schemes for multi-pass models and speaker adaptive training (assuming one video per speaker) are performed. The parameter tying for the different models is performed according to a decision tree using questions regarding SOU context defined on SOU clusters, as is done in [7].

The acoustic model training process does not require unit time marks, but rely on iterative alignment and model estimation. Only maximum likelihood training is performed for audio SOUs, as discriminative training may overfit to the noisy transcripts. With each SOU being a word, bigram and trigram language models are estimated using unannotated audio and the SOU models, also through decoding and estimation in an iterative fashion. We formulate the training process as follows:

$$\hat{\theta} = \arg\max_{\theta} \max_{W} p\left(X|W,\theta\right) p\left(W|\theta\right), \tag{1}$$

where θ is the union of acoustic model and language model parameters, W is the SOU label sequence, and X is the observed sequence of frame-level descriptors. Equation 1 is solved through iterative maximization, where in the i^{th} iteration, the following two steps are performed sequentially:

- Maximum likelihood training:

$$\theta_i = \arg\max_{\theta} p\left(X, W_{i-1}|\theta\right); \tag{2}$$

- Viterbi decoding: $W_i = \arg\max_W p\left(X, W|\theta_i\right)$.

We highlight three sets of models corresponding to the three passes in decoding. All these models are generated in the above training procedure. Please refer to previous speech recognition literature [8] for details about these models.

- The forward pass (FW) uses state-tied mixture models with non-cross-word triphones.
- The backward pass (BW) uses state-clustered-tied mixture models with non-cross-word triphones.
- The rescoring pass (RESC) uses state-clustered-tied mixture models with cross-word quinphones.

4. Unsupervised Audio SOU-based Representations

Decoding with audio SOUs and SOU language models is no different from ASR decoding, and is done in two stages. First, unadapted decoding (UDEC) is performed directly using trained

models, i.e. speaker independent. Second, in the adapted decoding (ADEC) stage, the SOU HMMs and the cepstral features are adapted according to the observed data for the specific video. In particular, three model-based MLLR adapations are done for the forward pass (FW) , the backward pass (BW) and the rescoring pass (RESC) respectively. Each of these MLLR adaptation can use a single transform matrix or multiple matrices according to a Gaussian clustering tree.

We extract various representations from the above multipass decoding process for high-level event detection.

4.1. Representations based on MLLR matrices

For HLDA-transformed cepstral features of F dimensions, each MLLR matrix is of size $F \times (F + 1)$ including the column for the bias. We devise a vector representation by stacking the matrix elements and perform L2 normalization. Three versions of such representations are produced, corresponding to the FW, BW, RESC passes in ADEC. Besides, for each pass we can consider the MLLR matrices only at the root node, only at the leaf nodes, or at all nodes of the Gaussian clustering tree.

Optionally, these representations from different passes can be stacked together as a single MLLR representation.

4.2. Ngram Representations from audio SOU lattices

The RESC step in ADEC produces the final decoding output. Instead of directly using its 1-best hypothesis, which is prune to noise, we leverage the SOU lattice output of the decoding process. We calculate the posteriors of SOU 1-grams, 2-grams, 3-grams and 4-grams from these lattices, aggregated within each video. A list of "adopted SOU ngrams" are identified for each of these according to a frequency threshold, or a TF-IDF threshold. These posteriors, with L2 normalization, constitute the vector representation for the audio SOU n-gram, $n \in \{1..4\}$.

5. Experiments

5.1. Dataset and setup

We tested our approach on a large, benchmark dataset of ∼45000 videos used in the TRECVID 2011 MED Evaluations [1]. The dataset contains ten complex events of interest described in Table 1, along with a large collection of diverse background video instances. We removed 2046 videos with possibly corrupt audio signal, e.g., constant static over time, and extracted audio signal with 8k sampling frequency for this work. Audio information is complementary in this heterogeneous dataset that is obtained from Youtube videos with almost unconstrained domains. All experiments in this work only uses the audio information, adding to the challenges of the task.

EventID	Description	# Training	# Evaluation
E006	birthday party	182	186
E007	changing a vehicle tire	112	111
E008	flash mob gathering	174	132
E009	getting a vehicle unstuck	128	95
E010	grooming an animal	137	87
E011	making a sandwich	125	140
E012	parade	157	231
E013	parkour	113	104
E014	repairing an appliance	130	78
E015	working on a sewing project	129	81
NULL	(background)	11695	30576
Total		13081	31821

Table 1: TRECVID MED11 dataset statistics

The detection of all the ten target events are carried out separately accordingly to the TREDVID MED 2011 guidelines. For each event, a binary linear classifier [11] is trained to distinguish between clips with target events and other clips. The hyperparameters of the SVMs are estimated through cross-validation grid search within the training set, for each representation separately. These settings make all the systems reported eligible audio components of the TRECVID MED evaluation.

A contrastive system using highly optimized soft quantization histograms based on Mel frequency cepstral coefficients (MFCC) is also included [3, 12], and is fused with the best audio SOU based classifiers using a fusion approach averaging the classifier output scores [3].

5.2. Local audio descriptor

Perceptual Linear Predictive (PLP) coefficients [13] are extracted from 25 ms windows at 100 frames per second. Each frame is windowed with a Blackman window and the power spectrum is computed for 125-3750 Hz from the signal sampled at 8kHz. From this, 30 Mel-warped spectral coefficients are computed. 14 PLP coefficients are calculated for each frame. At a given time step, a HLDA transform observes 15 consecutive frames of PLP coefficients, and produces a 60-dimensional descriptor.

5.3. Audio SOU based representations

In these experiments, we specify the number of SOUs to be 64, each modeled by a five-state left-to-right HMM. These SOUs are trained according to methods discussed in Section 3.

For the ngram representations derived from the audio SOU lattices, we consider the **1-gram**, **2-gram**, **3-gram** and **4-gram** as discussed in Section 4.2. For the **2-gram**, **3-gram** and **4-gram** representations, we only include audio SOU ngrams that occur at least 40 times in the training data. An additional representation **Ngram** concatenating all the four SOU ngram representations above is also considered.

For the representations derived from the MLLR transform matrices, we consider the following as discussed in Section 4.1. Only the matrices at the root of the Gaussian clustering trees are used in these experiments.

- **MLLR-FW**: only using the transform matrix from forward pass in decoding.
- **MLLR-FW-BW**: using the transform matrices in both forward pass and backward pass.
- **MLLR-all**: using the transform matrices from all passes.

An additional representation **SOU** combines **Ngram** and **MLLR-all**.

5.4. Experiment results

Based on event detection performance in Table 2, we observe the following about audio SOU based representations:

- Among the four ngram posterior representations from the audio SOU lattices, longer audio SOU ngrams encode more contextual information. Representations based only on the longer SOU ngrams do not always lead to improved performance.
- **Ngram** outperforms representations based on each individual type of audio SOU ngrams.
- Among the MLLR representations, using matrices from all three passes leads to the highest performance.

Table 2: Event detection performances (AUC) across ten events in Table 1.

Feature	Dim	Mean	E006	E007	E008	E009	E010	E011	E012	E013	E014	E015
1-gram	64	0.723	0.809	0.665	0.746	0.717	0.602	0.722	0.759	0.636	0.873	0.700
2-gram	3,552	0.748	0.846	0.690	0.793	0.748	0.627	0.708	0.774	0.695	0.893	0.710
3-gram	16,094	0.743	0.837	0.679	0.797	0.754	0.602	0.701	0.770	0.688	0.895	0.703
4-gram	15,302	0.720	0.809	0.657	0.781	0.728	0.580	0.680	0.726	0.665	0.888	0.687
Ngram	35,012	0.764	0.851	0.730	0.859	0.746	0.629	0.727	0.780	0.686	0.907	0.721
MLLR-FW	3,660	0.764	0.845	0.675	0.867	0.733	0.673	0.724	0.767	0.757	0.866	0.733
MLLR-FW-BW	7,320	0.772	0.854	0.661	0.872	0.769	0.694	0.740	0.779	0.752	0.880	0.722
MLLR-all	10,980	0.777	0.857	0.680	0.875	0.765	0.698	0.741	0.788	0.756	0.888	0.722
SOU	45,992	0.787	0.865	0.702	0.881	0.777	0.707	0.748	0.802	0.755	0.899	0.737
MFCCsoft	5,000	0.802	0.853	0.683	0.867	0.836	0.731	0.723	0.838	0.807	0.899	0.791
MLLR-all+MFCCsoft	fused	**0.810**	0.876	0.702	0.894	0.838	0.759	0.734	0.826	0.804	0.902	0.766
SOU+MFCCsoft	fused	**0.812**	0.876	0.716	0.897	0.839	0.758	0.737	0.828	0.784	0.912	0.777

- **SOU** is the best performing audio SOU representation, using all information from the SOU decoding process.

We notice that the audio SOU based representations do not yet outperform soft quantization histograms [12] of MFCCs in the contrastive system [3]. This is not surprising for two reasons. First, the contrastive system was highly optimized for the event detection problem in this dataset, while many parameters in the audio SOU training and decoding were optimized for speech processing. Second, it is observed in previous audio content analysis works [14] that the frame-level bag-of-words based methods outperform alternative methods based on higher level representations. However, **SOU** and **MFCCsoft** capture different information according to Table 2, demonstrated by the underscored AUC values marking the better between them. Systems combining the two produce superior results, as shown by the performance of the **MLLR-all+MFCCsoft** and **SOU+MFCCsoft** systems, which simply averages the classifier outputs in **MFCCsoft** and the respective audio SOU classifiers.

Besides achieving promising high-level event detection performance, the audio SOU approach produces an audio SOU lattice for each video, which is a rich and compact abstraction from the heterogeneous audio signal that can be used to produce different levels of segmental information. To verify that the audio SOUs indeed capture such information at a level higher than frame-based descriptors, we identify the most frequent audio SOU ngrams with length up to four, and calculate their durations from the audio SOU lattices in the training set. Only ngrams with lattice posterior higher than 0.5 are considered. In Table 3 we can see that these audio SOU ngrams represent much longer temporal structures than frame-based (25ms windows at 10ms time steps) audio descriptors used in the contrastive MFCC soft quantization histograms. There are SOU 3-grams and 4-grams that consist of the same consecutive audio SOUs. Preliminary examination suggests these units capture the acoustic background information.

Table 3: Duration statistics for most frequent SOU ngrams (in seconds, $\pm$ standard deviation)

SOUs	Duration	SOU 2-grams	Duration
P50	0.216 ± 0.209	P38-P38	1.055 ± 0.859
P21	0.197 ± 0.155	P13-P13	1.318 ± 1.157
P51	0.198 ± 0.189	P41-P41	1.483 ± 1.320
P13	0.619 ± 0.688	P50-P28	0.351 ± 0.245
P38	0.534 ± 0.549	P32-P48	0.489 ± 0.392
P53	0.199 ± 0.150	P50-P21	0.364 ± 0.250
P30	0.221 ± 0.155	P28-P50	0.358 ± 0.200
P17	0.216 ± 0.148	P50-P17	0.387 ± 0.246
P52	0.257 ± 0.206	P50-P30	0.415 ± 0.251
P32	0.243 ± 0.242	P39-P39	0.890 ± 0.716
SOU 3-grams	Duration	SOU 4-grams	Duration
P38-P38-P38	1.607 ± 1.126	P38-P38-P38-P38	2.313 ± 1.408
P13-P13-P13	2.167 ± 1.651	P13-P13-P13-P13	3.153 ± 2.067
P41-P41-P41	2.199 ± 1.629	P41-P41-P41-P41	3.111 ± 1.929
P39-P39-P39	1.345 ± 0.939	P13-P12-P13-P13	1.893 ± 1.263
P13-P12-P13	1.507 ± 1.413	P13-P13-P12-P13	1.982 ± 1.427
P50-P28-P50	0.483 ± 0.262	P39-P39-P39-P39	1.918 ± 1.144
P33-P33-P33	1.291 ± 0.992	P33-P33-P33-P33	1.743 ± 1.164
P48-P32-P48	0.781 ± 0.573	P38-P38-P38-P50	1.584 ± 0.971
P13-P36-P13	2.286 ± 1.990	P32-P48-P32-P48	0.963 ± 0.508
P32-P48-P32	0.778 ± 0.558	P39-P38-P39-P38	2.204 ± 1.487

6. Conclusions and Discussion

We identify the challenges of high-level event detection in heterogeneous data, and adapt the BBN SOU approach developed for speech processing for this application in audio content analysis. The proposed approach captures mid-level segmental information with no need of predefined mid-level concepts and annotated training data. The rich information in the multi-pass audio SOU decoding process enables the design of various audio representation.

We performed event detection on the benchmark TRECVID 2011 Multimedia Event Detection dataset using the proposed audio SOU based representations. Two sets of representations are evaluated in our experiments, generated using ngram posteriors estimated from the audio SOU lattices, and the MLLR transform matrices in model adaptation respectively. The proposed audio SOU approach not only achieves promising multimedia event detection performance, but captures segmental information at a level much higher than frame-level descriptors, lending the potential for human interpretable representation.

We plan to investigate further in three directions. First, the current audio SOU implementation has many parameters in model training and decoding that were optimized for speech processing. We will optimize them for audio processing. Second, the front end signal processing can be further improved by looking at a wider frequency range and different feature extraction that are more suitable for general audio analysis than human speech analysis. Third, we would like to quantitatively evaluate the correlation between the audio SOUs and various units used in previous work for multimedia event detection, including speech keywords, short (25 ms) frames and salient localized events. Alternative classifiers other than SVM may also provide more interpretable results.

7. References

[1] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, and G. Quenot, "Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2011*. NIST, USA, 2011.

[2] M. A. et. al., "The 2012 SESAME Multimedia Event Detection (MED) and Multimedia Event Recounting (MER) Systems," in *Proceedings of NIST TRECVID Workshop*, MD, United States, 2012.

[3] P. Natarajan, P. Natarajan, V. Manohar, S. Wu, S. Tsakalidis, S. N. Vitaladevuni, X. Zhuang, R. Prasad, G. Ye, D. Liu, I.-H. Jhuo, S.-F. Chang, H. Izadinia, I. Saleemi, M. Shah, B. White, T. Yeh, and L. Davis, "Bbn viser trecvid 2011 multimedia event detection system," in *Proceedings of NIST TrecVid 2011 Workshop*, Gaithersburg, MD., 12 2011.

[4] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Highlight sound effects detection in audio stream," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, vol. 3, 2003, pp. III–37–40 vol.3.

[5] M. Baillie and J. M. Jose, "Audio-based event detection for sports video," in *Proceedings of the 2nd international conference on Image and video retrieval*, ser. CIVR'03. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 300–309. [Online]. Available: http://dl.acm.org/citation.cfm?id=1760167.1760206

[6] K. Lee and D. P. W. Ellis, "Audio-based semantic concept classification for consumer video," *Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 6, pp. 1406–1416, Aug. 2010. [Online]. Available: http://dx.doi.org/10.1109/TASL.2009.2034776

[7] M.-H. Siu, H. Gish, A. Chan, and W. Belfield, "Improved topic classification and keyword discovery using an HMM-based speech recognizer trained without supervision," in *INTERSPEECH*, 2010, pp. 2838–2841.

[8] S. Matsoukas, J. L. Gauvain, G. Adda, T. Colthurst, C.-L. Kao, O. Kimball, L. Lamel, F. Lefevre, J. Z. Ma, J. Makhoul, L. Nguyen, R. Prasad, R. Schwartz, H. Schwenk, and B. Xiang, "Advances in transcription of broadcast news and conversational telephone speech within the combined ears bbn/limsi system," *Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 5, pp. 1541–1556, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1109/TASL.2006.878257

[9] M.-H. Siu, H. Gish, S. Lowe, and A. Chan, "Unsupervised audio patterns discovery using hmm-based self-organized units," in *INTERSPEECH*, 2011, pp. 2333–2336.

[10] S. Chaudhuri, M. Harvilla, and B. Raj, "Unsupervised learning of acoustic unit descriptors for audio content representation and classification," in *INTERSPEECH*, 2011, pp. 2265–2268.

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[12] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[13] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990. [Online]. Available: http://link.aip.org/link/?JAS/87/1738/1

[14] B. L., Y. SI, L. ZZ, O. A., J. Q., L. B., G. M., B. S., M. F., and H. A., "Informedia@TRECVID 2011," in *Proceedings of NIST TRECVID Workshop*, MD, United States, 2011.