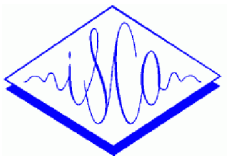

First Workshop on Speech, Language and Audio in Multimedia
Marseille, France
August 22-23, 2013

Audio Concept Ranking for Video Event Detection on User-Generated Content

Benjamin Elizalde[1], *Mirco Ravanelli*[2], *Gerald Friedland*[1]

[1]International Computer Science Institute, 1947 Center Street,
Berkeley, CA 94704, USA
[2]Fondazione Bruno Kessler, via Sommarive 18, 38123 Trento, Italy

benmael@icsi.berkeley.edu, mravanelli@fbk.eu, fractor@icsi.berkeley.edu

Abstract

Video event detection on user-generated content (UGC) aims to find videos that show an observable event such as a wedding ceremony or birthday party rather than an object, such as a wedding dress, or an audio concept, such as music, speech or clapping. Different events are better described by different concepts. Therefore, proper audio concept classification enhances the search for acoustic cues in this challenge. However, audio concepts for training are typically chosen and annotated by humans and are not necessarily relevant to a specific event or the distinguishing factor for a particular event. A typical ad-hoc annotation process ignores the complex characteristics of UGC audio, such as concept ambiguities, overlap, and duration. This paper presents a methodology to rank audio concepts based on relevance to the events and contribution to the ability to discriminate. A ranking measure guides an automatic selection of concepts in order to improve audio concept classification with the goal to improve video event detection. The ranking aids to determine and select the most relevant concepts for each event, to discard meaningless concepts, and to combine ambiguous sounds to enhance a concept, thereby suggesting a focus for annotation and a better understanding of the UGC audio. Experiments show an improvement of the audio concepts mean classification accuracy per frame as well as a better-defined diagonal in the confusion matrix and a higher relevance score. In terms of accuracy, the selection of top 40 audio concepts using our methodology outperforms the highest-accuracy-based selection by a relative 17.56% and a frame-frequency-based selection by 5.74%. In terms of relevance to the events, the ranking-based selection provided the highest score.
Index Terms: event detection, audio concept, user generated content, acoustic video processing.

1. Introduction

Video event detection aims to identify videos with a semantically defined event, such as a marriage proposal. This task is implicitly multimodal because events are characterized by audio-visual cues. Multimedia detection has been explored by computer vision using different features and techniques. However, audio has been under-explored, and state-of-the-art audio-based techniques do not yet provide significant assistance to its video counterpart. Audio, however, can sometimes be more descriptive than video, especially when it comes to the descriptiveness of an event. For instance, the audio cue can quickly allow one to determine whether or not a marriage proposal was successful. Thus, there is great importance in exploring techniques to improve the use of audio for video event detection.

There have been several approaches to audio-based video event detection for UGC data. Approaches in general employ only low-level features [1] [2]. However, there are also higher-level approaches that employ audio concepts for video event detection, motivated by the idea that different events are better described by different concepts. There are techniques that automatically derive audio concepts. An example is a system [3] which extracts audio units automatically with a diarization system to create an audio concept vocabulary. A similar example is a system in [4] that defines an automatic audio concept vocabulary with a Random Forest (RF) algorithm. However, these abstract representations may or may not map to a specific humanly understandable sound, such as clapping or the buzzing of a power tool. An example of an approach with annotated audio concepts for video event detection is [5] [6]. Whether these concepts are abstract or not, they define an acoustic fingerprint that distinguishes an event from their cohorts. The relation of concepts and events can be exemplified with a language analogy as stated in [3] where concepts can be seen as words and events as ideas. The paper shows that events are defined by different distributions of concepts. Therefore, improving the classification performance of concepts enhances the detection performance of events. Following this research line, the paper [7] aims to improve audio concept classification on UGC.

Nowadays UGC videos can provide massive amounts of training data, because the videos are widely available. Ad-hoc annotations of audio concepts for video event detection on UGC videos present three main issues. One is to ignore the intrinsic characteristics of UGC, where a concept could be in the presence of background noise, be overlapped with one or more concepts, have a short duration, be unintelligible for the annotator and have acoustic ambiguities with other concepts. The second is that audio concepts for training are typically chosen and annotated by humans and are not necessarily relevant to a specific event or the distinguishing factor for a particular event. The last issue lies in the performance of the audio concept classification by the technology employed. Adding audio concept annotations alone do not help as much as in other tasks such as speech detection, where in general the more annotated speech the better the detection performance. Take for instance a set of audio concepts that can be classified with high accuracies; if the concepts are not relevant to the events, they will be of little help to discriminate between events. On the other hand, let's assume we have a relevant and unique set of an event's audio concepts, which are not classified with reliable accuracies, then the concepts would be of little help to show evidence of the event detection. Therefore, the need to define a selection procedure that addresses the issues is presented in order to maximize the usage of current au-

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

Table 1: There are audio concepts annotations of at least 10 videos from each event.

Code	Event
E001	Attempting a board trick
E002	Feeding an animal
E003	Landing a fish
E004	Wedding ceremony
E005	Working on a woodworking project
E006	Birthday party
E007	Changing a vehicle tire
E008	Flashmob gathering
E009	Getting a vehicle unstuck
E010	Grooming an animal
E011	Making a sandwich
E012	Parade
E013	Parkour
E014	Repairing an appliance
E015	Working on a sewing project

dio concept annotations and understand the UGC audio better.

This paper presents a methodology to rank audio concepts based on relevance to the events and contribution to the ability to discriminate. The ranking guides an automatic or user-based selection of concepts in order to improve audio concept classification for video event detection. The ranking aids to determine and select the most relevant concepts for each event, discard meaningless concepts, combine ambiguous sounds to enhance a concept, thereby suggesting a focus for annotations. The paper also provides an analysis on the UGC audio concept annotations.

The content of the paper is structured as follows. Section 2 presents the UGC video and the audio concept annotations for the experiments. Section 3 details the ranking methodology. Section 4 describes the audio concept classification system and the experiments. Section 5 continues with the results and expands the understanding of the UGC audio characteristics. Lastly, Section 6 states the conclusion and future work.

2. UGC Video and Annotations Sets

The video set used for the audio concept annotations is the NIST TRECVID MED 2012, which contains UGC videos. The 2012 corpus consists of 150,000 videos of about three minutes each. The audio from the videos contains environmental acoustics, overlapped sounds, and unintelligible audio among other characteristics. The annotations are based on the Event Kits subset. Table 1 contains a summary of the events.

The annotation set from SRI-Sarnoff consists of manually labeled sounds of 291 videos. The videos belong to the 15 events of the MED 2012 Event Kits dataset for a total of 11.6 hours. In total there are 28 audio concepts shown in Table 2, which attempt to describe distinctively the events.

The annotation set from CMU [8] consists of manually labeled environmental acoustics of 216 videos taken from MED 2012, totaling 5.6 hours There are at least 10 annotated videos for each of the 15 events from MED 2012. The result is a set of 42 audio concepts shown in Table 3. The main goal of the annotations was to create labels for audio segments that exist solely in the audio domain.

Table 2: List of 28 audio concepts annotated by SRI-Sarnoff in alphabetical order.

1	audio of wedding vows	15	instructional speech
2	bagpipes	16	landing after a jump
3	blowing out candles on a cake	17	laughing
4	board hitting surface	18	marching band
5	cheering	19	metallic clanking noises
6	childrens voices	20	music
7	clapping	21	noise of passing cars
8	clinking	22	power tool whine
9	conversational speech	23	rolling
10	crowd noise	24	sewing machine sound
11	dancing singing in unison in a group	25	singing
12	drums	26	someone giving a speech
13	group dancing	27	word how spoken
14	group walking	28	word tire spoken

3. Ranking Methodology

The ranking methodology is an iterative process that is divided in four steps. The first step is to calculate the relevance of the audio concepts based on how rare or common the concepts are to a specific event and to the rest of the events. The second step is to run our Audio Concept Classification system and measure the classification performance for each concept. The third step is to calculate the ranking of the audio concepts by considering the results from step one and two for each concept. Finally the fourth step consists of deciding whether a concept should be merged with another or discarded. The process iterates until the desired final quantity of concepts is reached.

3.1. Step 1: Compute relevance

The relevance of a concept to an event is expressed by the well known algorithm of Term Frequency - Inverse Document Frequency (TF-IDF) [9]. The raw frequency is the number of times a term*t* occurs in a specific document *d*. To prevent a bias with unbalanced documents, the raw frequency is divided by the maximum raw frequency of any term in the document. The TF is defined by the equation 1.

$$TF(t,d) = \frac{f(t,d)}{max\{f(w,d) : w \in d\}} \quad (1)$$

The IDF tells you whether a word is common or rare across the documents. It is the result of taking the logarithm from the division of the total number of documents by the number of documents containing the term. If the term is not in the corpus, the division will be zero, thus we add *1*. The IDF can be defined by the equation 2.

$$IDF(c,D) = log\frac{|D|}{1 + |\{d \in D : c \in D\}|} \quad (2)$$

A high TF-IDF score is reached by a high term frequency in the given document and a low document frequency of the term in the whole collection of documents; the scores therefore tend to filter out common terms. In our methodology a term

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

Table 3: List of 42 audio concepts annotated by CMU in alphabetical order.

1	anim bird	15	engine light	29	rustle
2	anim cat	16	engine quiet	30	scratch
3	anim goat	17	hammer	31	scream
4	anim horse	18	human noise	32	singing
5	applause	19	knock	33	speech
6	bang	20	laugh	34	speech not english
7	beep	21	micro blow	35	squeak
8	cheer	22	mumble	36	thud
9	child	23	music-sing	37	tone
10	clap	24	music	38	washboard
11	clatter	25	phone	39	water
12	click	26	power tool	40	whistle
13	crowd	27	processed	41	white noise
14	engine heavy	28	radio	42	wind

corresponds to a frame from an audio concept and a video event category corresponds to a document.

3.2. Step 2: Measure performance

The classification performance (CP) of the technology employed should be considered to let the system decide which concepts are more meaningful and distinguishable, along with the limitations of using a determined audio concepts set. In this paper a raw classification accuracy per frame metric is included in the ranking equation 3, but it could be substituted by any other metric that evaluates the concept CP. In this step the confusion matrix for the list of concepts is also computed in order to determine the confusability of each concept in respect to the others.

3.3. Step 3: Compute ranking

The ranking represents the relevance and classification performance for each audio concept. A higher *Rank* score means more relevance to the events and it can be represented by equation 3. The ranking is a single score for each audio concept that consists of multiplying the TF, times the IDF, times the CP across the events.

$$Rank(c, D) = (TF(c, d) \cdot IDF(c, D) \cdot CP(c)) \quad (3)$$

3.4. Step 4: Merge or discard

Once the ranking scores are computed for each audio concept comes the decision as to whether to merge or discard the lowest ranked concept. The lowest ranked concept *C-low* would be merged with the corresponding most confusable concept *C-conf* according to the confusion matrix. The cohort concepts are merged because *C-low* has low relevance and is not discriminating and distinguishable enough. The concept with higher relevance and accuracy that absorbed *C-low* will provide the name to the new resulting concept *C-merged* and will keep its corresponding annotation data. The audio classification system is run again and if the classification accuracy of *C-merged* increases, then it remains as it is. The ranking process continues the next iteration removing *C-low* from the list, but keeping its annotation data. In case the accuracy of the *C-merged* did not increase, then *C-low* is not merged and instead is discarded from the list along with its annotation data. Once again the process continues with its next iteration until the desired number of concepts is reached.

4. Experimental Setup

This section describes the classification system and details the most relevant experiments.

4.1. Audio Concept Classification System

The audio concept classification system is based on a Neural Network approach because it has demonstrated high performance on a similar task where it discriminates well between different sounds called phonemes [10] [11]. The system employs the Parallel Neural Network Trainer TNet [12] technology from Brno University of Technology. The Neural Network (NN) architecture is basic and is the first step to move on to Deep Learning, it consists of two hidden layers with 1,000 neurons each and sigmoid activation functions. The extracted acoustic features are the typical Mel-Frequency Cepstral Coefficients (MFCCs) C0-C12, with energy included, for a total of 13 dimensions. Each feature frame is computed using a 25 ms hamming window, with 10 ms frame shifts. The neural network was fed, after a mean and variance normalization step, by the specified features using a context window of nine consecutive frames. The output layer, whose softmax-based neurons dimensionality is equal to the number of audio concepts to classify. More specifically, for the training phase a stochastic gradient descent optimizing cross-entropy loss function was used. The learning rate was updated by the "newbob" algorithm: It is kept fixed at LR=0.002 as long as the single epoch increment in cross-validation frame accuracy is higher than 0.5%. For the subsequent epochs, the learning rate is being halved until the cross-validation increment of the accuracy is inferior to the stopping threshold 0.1%. The NN weights and biases are randomly initialized and updates were performed per blocks of 1024 frames.

4.2. Experiments

The objective of the experiments consists in selecting the top 40 audio concepts that provide the best trade-off between classification performance and relevance to the events. The reason for choosing 40 is that out of the 70, this is the largest number of concepts that our system was able to classify with more than one percent of accuracy. The first experiment consists of using a concept set from a selection based on highest-accuracy. The 70 concepts are fed into the audio concept classification system and then sorted to select the top 40 with highest classification accuracy. The reason for this selection is because intuitively it will lead to a high overall concepts accuracy. The second experiment uses a set based on a high frame-frequency selection. The 70 concepts annotations are analyzed and then the concepts are sorted based on the quantity of frames. The selection is motivated because concepts with more frames will most likely have longer durations or be more common, which makes them easier for the system to classify them, and more important they will have more training data available. Lastly the third experiment employs a selection based on the ranking presented in this paper.

The annotations add up to 17.2 hours and are separated into training and test. The training set contains 90% of the annotations for a total of 15.48 hours. The test set consists on the

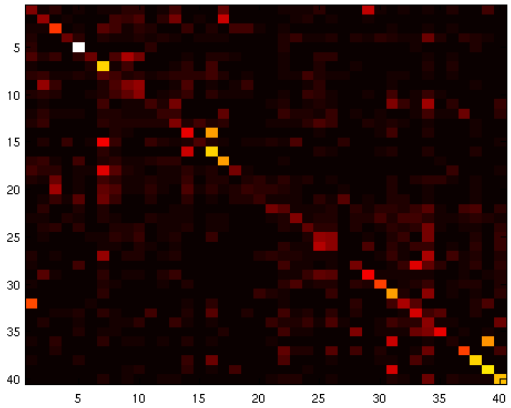

Figure 1: *The confusion matrix based on the top 40 highest-accuracy set, shows dark regions in the center of the diagonal.*

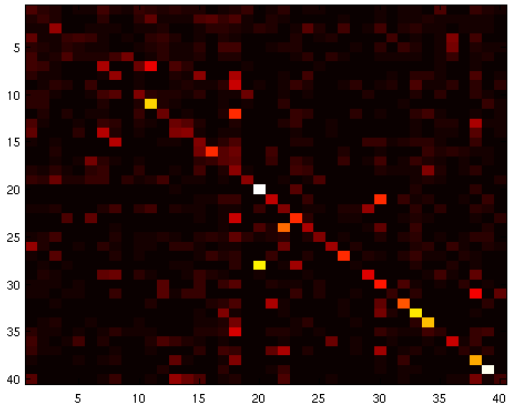

Figure 2: *The confusion matrix based on the top 40 frame-frequency set, shows dark regions on the up-left part of the diagonal.*

other 10% for a total of 1.72 hours. The classification accuracy per frame is evaluated by comparing the label from the frame's highest posterior against its corresponding label from the ground truth.

In order to provide a baseline for the three experiments against using the total number of concepts, the audio concept classification system is trained with the 70 concepts. The overall mean classification accuracy per frame is 11.5 % with a random guess of 1.42%.

5. Results and Analysis

This section presents the results from the experiments and expands the understanding of the UGC audio characteristics derived from our experiments and an analysis of both concept annotation sets.

5.1. Results

The classification performance of the three experiments is shown in the second column of Table 4. The first experiment with a highest-accuracy-based selection has an overall mean accuracy per frame of 20.38%, while the second experiment with a frame-frequency-based selection has 18.33%. The third experiment using the ranking-based set shows 21.55%. The selection of the top 40 audio concepts using our methodology outperforms the highest-accuracy-based selection by a relative 17.56% and the frame-frequency-based by a relative 5.74%.

The level of relevance to the events of the three sets of audio concepts is shown in the third column of Table 4. The score for each set is the normalized log TF-IDF, which consisted of three steps: First, the TF-IDF scores for the 40 concepts are computed as in steps one and two from Section 3. Second, the log of the TF-IDF score is computed. Finally, on the third step, a normalization is applied to the three scores, where the highest possible value of the three sets equals to one, and the other two are proportional. The lower the value, the lower the overall relevance of the set to the 15 events. The log and the normalization steps are meant to provide a more human understandable comparison. The ranking-based selection provided the highest relevance score for the 15 events, with an improvement of 17% in respect to the highest-accuracy set and 10% in respect to the frame-frequency set.

The confusion matrix is a table that allows the visualization of the accuracy performance of the system, in other words, it shows the confusability of each concept in respect to the others. Each column of the matrix represents the instances of the predicted concept, while each row represents the instances of the actual concept. A better defined diagonal means less ambiguities and higher accuracy, hence a more distinguishable and distinctive set of concepts.

The confusion matrix of the highest-accuracy set experiment is shown in Figure 1, the frame-frequency-based matrix is in Figure 2 and the one from the ranking-based set is in Figure 3.

In terms of usage of the annotated data, the 70 audio concepts comprehend about 683 minutes. The frame-frequency-based selection used 664 minutes, while the highest-accuracy selection uses 577 minutes and the ranking-based used 668 minutes. Our approach uses slightly more minutes or frames than the highest-accuracy set, which means that most of the information of the annotations is been used.

The results confirm that our methodology provided the best overall classification accuracy, the least concepts confusability and the best relevance of the audio concepts to the 15 events.

5.2. Analysis of the audio concepts

This section intends to aid the understanding of the UGC audio. The following includes an analysis of the annotation sets regarding concept overlap and duration. In addition, there is an analysis of the concepts merging and discarding step from our methodology to explain concept ambiguity.

The video events are described by a set of different sounds that occur throughout the recording therefore making it possible that one or more concepts occur at the same time, resulting in an overlap. The annotations has 38% of audio overlapping with one or more concept. The most common types of overlap are music and other audio concepts except speech 35% of the time, speech and other concepts except music 13% and speech and music 4%. The rest of the overlap types complete the total with 48%. The situation of having three or more annotated overlaps is rare and it accounts for less than 3% of the audio. It is important to mention that there could be other concepts that overlap

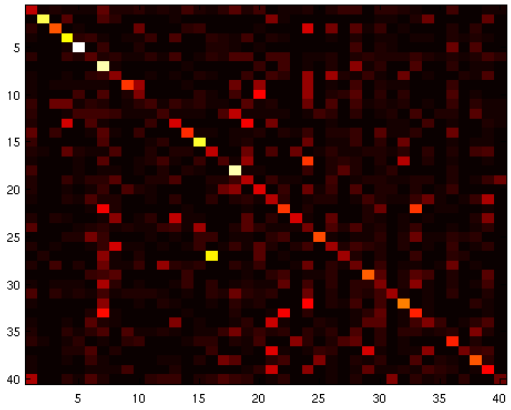

Figure 3: *The confusion matrix based on the top 40 ranked set, shows a better defined diagonal than the other two sets.*

Table 4: The ranking-based selection shows the best overall mean accuracy and the highest relevance for the top 40 audio concepts.

Selection based on	Mean accuracy	TF-IDF score
Highest-accuracy	20.38 %	0.83
Frame-frequency	18.33 %	0.90
Ranking	21.55 %	1

and are not annotated.

The nature of the concepts in the events and their duration is diverse. The annotations show that the average duration of the trials or segments is about one second. For speech trials, 38% lasts less or equal than one second and 30% lasts less or equal than two seconds, but more than one second. Examples of audio concepts with long duration are music with trials of up to 380 seconds or crowd noise of up to 260 seconds. Examples of short duration concepts are beep with trials as short as 0.8 seconds or clinking as short as 0.5 seconds. In [8] indicated that for the CMU annotations, shorter trial durations have lower accuracies, and longer duration trials have better accuracies, which was confirmed in this research with the inclusion of the SRI-Sarnoff annotations. For example, music was detected with about 90% accuracy and crowd noise with 40%, while beep and clinking resulted in less than 2% accuracy.

Discarding concepts alone intuitively suggests an improvement in accuracy. Depending on the selection process, different audio concepts will be discarded, thus affecting the overall classification in different ways. Our iterative procedure does not discard concepts for the sake of them, instead it could merge audio concepts, resulting in a more analytical usage of the annotated audio. Experiments one and two discarded the lowest 30 concepts according to their selection type. The iterative procedure from the third experiment had 30 iterations, discarded 16 concepts and merged 14. Examples of concepts discarded are: blowing out candles on a cake (SRI), clinking (SRI), dancing singing in unison in a group (SRI).

Both of the annotation sets have unique characteristics, focus and annotators. Hence, even though some of the concepts have the same or similar logical name there is no reason to assume that they should be considered as the same concept. In our methodology, merging redundant concepts from different annotation sets could sometimes make sense to the user such as cheer (CMU) and cheering (SRI) or laugh (CMU) and laughing (SRI). Nevertheless, there are other situations where is not as logical to merge sounds. Audio concepts sometimes overlap and one of them may have more "prominent" acoustic characteristics than others such as volume, pitch, duration, etc. Take for instance, the concepts group dancing (SRI) and music (SRI). The first concept is overwhelmed by music (sometimes added by the user), which has higher prominence, thus significantly decreasing the classification accuracy of the concept. The overlap information can be extracted from the annotations, but not the prominence level of the concepts involved. More complicated is when merged sounds do not have a logical semantic relation, but they could make sense from the audio concept classification system perspective. Examples are squeak (CMU) and white noise, which are broadband sounds, or thud (CMU) and click (CMU), which are impulsive sounds, or animal cat (CMU) and scream (CMU), which have similar pitch. As part of the evolution of our work we would like to include user-intervention as prior information to figure out its impact on the results of the merging process. We understand that technology and events can change and whenever this happens the iterative ranking process could be re-applied using the original set of annotations.

6. Conclusions

The research shows that the ranking methodology aids the selection of audio concepts with the best trade-off between relevance to the event and classification accuracy. The methodology discards less relevant and less accurately detected concepts and merges ambiguous sounds to enhance a concept. More important is that the ranking serves to maximize the usage of current sound concepts annotations. The improvement in classification accuracy improves the classification of concepts which provide a more reliable evidence for video event detection. The selection of top 40 audio concepts using our methodology outperforms a highest-accuracy-based selection by a relative 17.56% and a frame-frequency-based selection by 5.74%. In terms of relevance to the events, the rank-based selection provided the highest relevance score, with 17% more than the highest-accuracy-based selection and 10% more than the frame-frequency-based selection. Furthermore, the ranking suggests the audio concepts that can be enhance by more annotations and the concepts that are less relevant to the technology. Future work involves using the classification posteriors output for video event detection. The output may be used for audio segmentation or as a semantic feature, both options can feed a video event detection system.

7. Acknowledgements

Thanks to Adam Janin for his advice. Thanks to Ajay Divakaran for the SRI-Sarnoff annotations.

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 Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusion contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

8. References

[1] R. Mertens, H. Lei, L. Gottlieb, G. Friedland, and A. Divakaran, "Acoustic Super Models for Large Scale Video Event Detection," in *ACM Multimedia*, 2011.

[2] X. Zhuang, S. Tsakalidis, S. Wu, P. Natarajan, R. Prasad, and P. Natarajan, "Compact audio representation for event detection in consumer media," in *INTERSPEECH*. ISCA, 2012.

[3] B. Elizalde, G. Friedland, H. Lei, and A. Divakaran, "There is No Data Like Less Data: Percepts for Video Concept Detection on Consumer-Produced Media," in *ACM International Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis at ACM Multimedia*, 2012.

[4] P.-S. Huang, R. Mertens, A. Divakaran, G. Friedland, and M. Hasegawa-Johnson, "How to put it into words - Using random forests to extract symbol level descriptions from audio content for concept detection," in *ICASSP*, 2012.

[5] Q. Jin, P. F. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based Video Retrieval Using Audio," in *Proceeding of the 13th Annual Conference of the International Speech Communication Association*, 2012.

[6] S. Pancoast, M. Akbacak, and M. Sanchez, "Supervised Acoustic Concept Extraction for Multimedia Event Detection," in *ACM International Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis at ACM Multimedia*, 2012.

[7] A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, and B. Raj, " Audio Event Detection from Acoustic Unit Occurrence Patterns ," in *ICASSP*, 2012, pp. 489 – 492.

[8] S. Burger, Q. Jin, P. F. Schulam, and F. Metze, " Noisemes: Manual Annotation of Environmental Noise in Audio Streams ," Tech. Rep., 2012.

[9] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for idf," *Journal of Documentation*, vol. 60, p. 2004, 2004.

[10] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[11] A. rahman Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[12] K. Vesely, L. Burget, and F. Grezl, "Parallel Training of Neural Networks for Speech Recognition," in *Proceeding of Interspeech*, 2010.