IMPROVED AUDIO FEATURES FOR LARGE-SCALE MULTIMEDIA EVENT DETECTION

Florian Metze, Shourabh Rawat, and Yipei Wang

Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA 15213; U.S.A.
{fmetze,srawat,yipeiw}@cs.cmu.edu

ABSTRACT

In this paper, we present recent experiments on using Artificial Neural Networks (ANNs), a new "delayed" approach to speech vs. non-speech segmentation, and extraction of large-scale pooling feature (LSPF) for detecting "events" within consumer videos, using the audio channel only. A "event" is defined to be a sequence of observations in a video, that can be directly observed or inferred. Ground truth is given by a semantic description of the event, and by a number of example videos. We describe and compare several algorithmic approaches, and report results on the 2013 TRECVID Multimedia Event Detection (MED) task, using arguably the largest such research set currently available. The presented system achieved the best results in most audio-only conditions.

While the overall finding is that MFCC features perform best, we find that ANN as well as LSP features provide complementary information at various levels of temporal resolution. This paper provides analysis of both low-level and high-level features, investigating their relative contributions to overall system performance.

Index Terms— acoustic event detection, computational acoustic scene analysis, multimedia retrieval

1. INTRODUCTION

Consumer-grade video material is becoming available in abundance over the Internet. Video analysis has progressed significantly since the early days of video retrieval, but significant progress is still needed before tasks such as TRECVID's Multimedia Event Detection (MED) and Recounting (MER) can be considered solved.

MED [1] consists of indexing and searching large sets of multimedia content, i.e. the HAVIC corpus [2], >5000 h, in order to retrieve from the collection the most relevant videos that contain instances of certain "events". Examples of such events include "Birthday Party" or "Repairing an Appliance", which are defined using textual descriptions (an "event kit") along with several example videos. Events have to be "observable" in either the visual or auditive domain, so it has been shown that the audio channel (i.e. automatic speech recognition and acoustic scene analysis) can contribute significantly to overall system performance, even though about a third of the videos do not contain useful audio at all, and only about a third of videos do contain useful English speech.

One of the most important trends in speech and language processing has been the introduction of Deep Learning (DL) [3], which has already lead to significant improvements in vision and image processing. In this paper, we will introduce DL to large-scale audio classification, and compare its use with other established techniques. As almost no labeled data is available for supervised training, we train sparse auto-encoders to derive complementary feature representations.

In audio processing, and MED in particular, one class of approaches makes use of a high-level semantic vocabulary [4, 5, 6, 7, 8], a set of predefined atomic acoustic event detectors learned in a supervised manner, on manually annotated data. Examples include sounds like "engine", "water", or even speech/ silence detection. Even though this approach provides semantically interpretable information about the event and instance, which is desirable in the context of video summarization (e.g. MER), it does not scale well to consumer videos, which are highly unstructured and unconstrained, and are generally expensive to annotate consistently.

Another approach is to learn completely unsupervised acoustic units directly from the data itself. A popular technique to learn these units is using a clustering algorithm such as k-Means. While these acoustic units may not have an attached semantic label, their distribution can convey information about the event, often referred to as the Bag-of-Audio-Words (BoAW) model, which has been applied successfully to many content-based audio retrieval tasks [9, 10, 11].

In a third, orthogonal line of research, Automatic Speech Recognition (ASR) can be used to transcribe spoken content in the audio track, assuming it contains enough speech to allow for text-based event detection [12]. Obviously, speech vs. non-speech segmentation can be challenging, as noise levels can be high, or speech can be overlaid with music or even other competing speakers. Also, in practice, data-sets can contain speech in many languages, potentially requiring multi-lingual recognition, and machine translation.

Finally, audio-visual approaches have been tried [13], in which joint learning of acoustic and visual models, either on the symbolic or purely data-driven level has been tried.

2. TRECVID MULTIMEDIA EVENT DETECTION

In this paper, we report results of experiments on the "MEDTEST" internal development dataset (ca. 837 h of video), and on the official "MED13" evaluation dataset (ca. 3,722 h) of the TRECVID 2013 Multimedia Event Detection (MED) task [1]. MED13 consists of the full set, and a "progress" sub-set.

The dataset comprises 38 events that have appeared in MED10 (P001-P003), MED11 (E001-E015), MED12 (E021-E030), and MED13 (E031-E040) evaluations. Events E031-E040 are known as the "ad-hoc" events, because they were not released to participants before the evaluation, and had to be processed roughly within a month of their disclosure, while the remaining events were known during development, and are therefore call "pre-specified" events. If the performance of a feature is significantly lower on "ad-hoc" events than on "pre-specified" events, this may therefore indicate over-specialization or some other mismatch.

We report results for three different conditions, EK 0, EK 10, and EK 100, which differ by the number of positive example videos for each class that are available for model training. E.g. in the EK 100 condition, 100 videos are available to estimate a model for each event (ca. 290 h in total), in addition to a semantic "event kit" description, while in EK 0 only the text of the event kit can be used, as no videos are available. The remaining videos are unrelated to the target events and provided as "background" videos.

For evaluation, we use the mAP (mean Average Precision) metric, which was used as a standard metric in the 2013 evaluation. It describes the mean precision achieved across all classes defined on the databases. For more details, please refer to the MED13 evaluation plan [14].

3. APPROACHES

One of the most successful strategies for building well-performing systems in video retrieval has been to fuse the output of several complementary features or classification approaches. In this section, we describe a number of such features, which were then combined using late fusion.

While a number of additional features and techniques were also tried during the preparation of our evaluation submission, none contributed significantly, so that we do not describe them in this paper.

Care has been taken to extract audio data consistently using FFmpeg [15] from all the data, which has been provided in MP4 containers.

3.1. ASR – Automatic Speech Recognition

Because of the noisy nature of the HAVIC video data, we developed a new "delayed" approach to automatic segmentation for audio data. It is optimized for retrieval performance, rather than Word Error Rate (WER), which is the commonly used metric during ASR development, and avoids discarding useful information during retrieval due to hard segmentation decisions. This approach effectively performs a soft assignment of speech segments to speech and non-speech categories, thereby ensuring that ASR output can be used for retrieval even in segments that were not reliably classified as speech. The ASR itself was trained on 100 h of English Broadcast News (BN) and about the same amount of Meeting (RT-04) data, and was implemented using the Janus toolkit with the Ibis decoder [16]. The Word Error Rate is about 60 % in an internal test set (different from the test sets used here). The ASR uses Gaussian Mixture Models (GMMs) trained using Maximum Likelihood, and a 4-gram language model.

The following steps are performed:

- Run a rough ergodic GMM-HMM-based first-pass segmentation on MFCC (Mel-Frequency Cepstral Coefficients) features using silence, noise, music, and speech categories, geared towards high recall.
- Perform ASR on all non-silence segments, using per-segment Cepstral Mean Subtraction/ Cepstral Variance Normalization for noise robustness; the ASR's acoustic model contains dedicated models for noise and music to account for false positives in this step.
- Collect vocabulary and estimate the Idf (Inverse document frequency) value for each term in the vocabulary from the resulting transcriptions on the training data, after applying stop-word removal and stemming. To increase variability, transcriptions were generated using both Maximum Likelihood (Viterbi) and Minimum Bayes Risk (MBR) decoding. Calculate Tf-Idf (Term Frequency/ Inverse Document Frequency, [17]) values using the learned dictionary.
- Perform soft segmentation using the semantic audio features trained on the union of 42 [4], 28 [6], and 20 [7] concepts, using classifiers trained on the LSPF features described below, at a 0.1 s frame rate. Treating the three semantic concept vocabularies as one vocabulary of 90 words, compute corresponding Tf-Idf values. Semantic audio features are trained on LSPF features using a random forest classifier, on about 1,000 features after they have been ranked using Information Gain.
- Filter the ASR output in Tf-Idf representation by concatenating the BOW (Bag of Words) representations of ASR and semantic audio features. These semantic audio concepts act as a filter, providing a strong signal for labels such as laughter, childrens speech, vocal music, etc. [4] in areas where the ASR would normally produce no useful result because of high WER. By concatenating these complementary representations, greater robustness is achieved in the ASR system.

In additional experiments, we improved the transcription step of ASR system by including hybrid acoustic models using the state-of-the-art Deep Bottle-Neck Feature (DBNF) approach, as described in [18]. Although the WER was reduced by about 5 % absolute, this did not result in reduction of mAP, because the number of tokens provided by this ASR was significantly lower than the number of tokens generated by the conventional GMM based ASR.

3.2. MFCC – Mel-Frequency Cepstral Coefficients based Bag of Words Features

The baseline MFCC system is a fusion of three straightforward MFCC BoAW representations. Features were computed with a 10 ms step size, using 20 ms windows. Cepstral mean subtraction and variance normalization was performed on a per-video level, using a power-based detection of target audio, as is customary in speech recognition.

Stack-5: stacking 5 neighboring frames of 13-dimensional MFCC features and training 16 k cluster centers with k-Means. Various temporal contexts were explored, with 5 frames performing best on the development data.

Compressed-16k: A hierarchically clustered version of the 16k k-Means cluster centers to reduce the dimensionality to 4096. No temporal context was used to increase diversity between features.

k-Means-4k: training 4096 cluster centers on MFCC features without any temporal context or hierarchical clustering.

The computation and extraction of MFCC features follows the approach described in [9]. In computational audio analysis, plain MFCC features were proven to be a very hard baseline to beat, with this work not being an exception. Slight gains in mAP could be achieved by fusing together several classifiers trained on multiple, slight variants of MFCC features, which cover a temporal context of about 100 ms, but no further gains in mAP could be achieved in this setup by learning for example self-organized audio units with similar temporal extent from the data [19].

3.3. lMEL – log-MEL-based Sparse Coding Features

The initial integration of Deep Learning into ASR [20] has been achieved by transforming traditional features like MFCCs or lMELs (log-based Mel Spectral Coefficients, [21]) into a Bottle-Neck representation, on which GMMs were then trained. Unsupervised pre-training using Restricted Boltzmann Machines or Sparse Denoising Autoencoders has been shown to be effective at conditioning the parameters in deep networks to achieve better performance during supervised fine-tuning [22]. lMEL features are often preferred over MFCCs as inputs, because ANNs can learn further data transformations by themselves, and are not hindered by correlations in the dimensions as are GMMs.

In our implementation, these features are learned by training a single-layer sparse coder in an unsupervised manner over lMEL features. The features are trained over 7 frame stacked lMEL features (210 dimensions) as the input layer and a 60 dimensional output layer (30 lMEL features of the center frame plus deltas) with a sparsity factor of 0.02. Once the features are learned, encoding and classification is performed using the BoAW approach similar to the one used in case of MFCC. Model training has been performed on GPUs using straightforward adaptations of the Theano toolkit [23].

3.4. LSPF – Large-Scale Pooling Features

A weakness of the previously presented MFCC and lMEL features is that they capture very little temporal context and no local coherence of the signal, even though temporal stacking has been applied. We therefore attempt to extract features from a wider acoustic context, and characterize classes by looking at properties of the audio signal over frame sizes greater than the 10 ms order of magnitude.

We first extract a number of low-level descriptors, such as MFCC, PLP, LPC, Pitch, Loudness, Chroma, Formants, LSP, Signal Energy, Spectral Flux, etc. as well as certain functionals, such as Means, Extremes, Moments, Peaks, Percentiles, Onsets, Zero-Crossing, etc., and use these functionals, extracted on the above descriptors, as features [24].

In our implementation, a set of 6,373 features is being extracted over 2 sec segments every half second using OpenSMILE's [25] "ComParE" configuration. Feature selection is performed using an Information Gain criterion followed by Principal Component Analysis and Whitening to reduce dimensionality to 100 or 300. These features are used both directly in the k-Means framework (fusing the two feature streams withe different dimensionality) and for trainining semantic ("noiseme") concept detectors, which in turn are then used for segmentation, as described in Section 3.1. Parameters have been tuned on the internal development set.

4. EXPERIMENTS AND RESULTS

Experiments and results are summarized in the following Tables. Table 1 shows the results of the individual low-level features presented in Section 3 on the internal MEDTEST development set, using the same k-Means based classification pipeline for all features.

It can be seen that the individual MFCC features are the most powerful features, and that adding temporal context by simply stacking neighboring frames as well as adding a hierarchical clustering for a compressed representation improve performance over "plain" MFCC features.

Table 1. Feature Comparison (mAP) on internal MEDTEST development data, using SVM classifiers. The top rows show performance on the pre-specified events, the bottom rows show ad-hoc events (E031-E040). Overall performance depends strongly on the number of available training samples (EK 100 vs. EK 10), with stacked MFCC features performing best.

mAP	MFCC-kMeans-4k BoAW	MFCC-Compressed-16k BoAW	MFCC-Stack5 BoAW	LSPF 300 BoS	lMEL-Sparse BoAW
EK100	13.57	14.49	14.62	12.49	12.95
EK10	8.28	9.20	9.42	6.13	7.77
EK100	12.57	12.48	13.28	6.93	11.46
EK10	5.90	6.56	6.46	3.88	4.54

Table 2. mAP performance of different features on MEDTEST development data, using SVM and KR classifiers, for ad-hoc events. The benefit from using delayed segmentation as presented in Section 3.1 for ASR is evident.

mAP on MEDTEST	EK 10 (KR)	EK 10 (SVM)	EK 100 (KR)	EK 100 (SVM)
ASR	4.25	3.94	8.61	7.22
ASR (non-delayed segmentation)	2.39	2.79	5.07	4.50
lMEL	3.97	4.54	10.14	11.46
LSPF 100	2.82	3.32	5.98	6.59
LSPF 300	3.05	3.88	6.51	6.93
MFCC-Compressed-16k	6.66	6.56	12.29	12.48
MFCC-k-Means-4k	5.81	5.90	12.04	12.57
MFCC-Stack5	6.33	6.46	12.62	13.28

Table 2 shows that the choice of classifier (Support Vector Machine, SVM, vs. Kernel Regression, KR) is secondary, compared to the choice of feature. This table also shows the influence of the proposed "delayed" segmentation on ASR classification performance, and the performance breakdown of the individual LSPF features without fusion.

5. SYSTEM COMBINATION AND FUSION

Late fusion was used to combine individual audio systems together to improve overall performance. Fusion performance was optimized on MEDTEST data.

It is interesting to note that the MFCC feature provides the best performance in itself. A comparison between Table 1 and Table 3 shows the benefit of the combination of three MFCC features in itself to be 0.76 mAP. Still, LSPF features contribute significantly to the overall performance, because leaving them out from the audio fusion results in the greatest overall loss (0.85 mAP), underscoring their complementarity. lMEL features also contribute. We also integrated two other types of features, sub-band auto-correlation features (SBPCA) [26], and self-organized units (SOUs) [19], however these had the least influence on overall performance, despite them being potentially complementary as well.

Table 4 shows the official evaluation results. The presented audio and ASR systems were tuned for EK 100, and performed best for most conditions, with almost no tuning effort for the other setups. No low-level features could be used in the EK 0 condition, so only the semantic "noiseme" features, as described in Section 3.1, were used in the "Audio" case. Words of the ASR output and names of the semantic concepts that could be detected were mapped to the terms contained in the event kits in this case. It can be observed that semantic audio features and ASR become relatively more important for the low-resource conditions, where models for data-driven approaches become harder to train.

Table 3. Performance on MEDTEST (EK 100 condition) for individual features and fusion of audio-only features (no ASR), and when leaving out individual features from the fused audio result ("Leave-out-loss" column).

mAP	Feature	Leave-out-loss
Audio Fusion	17.04	-
MFCC	15.38	0.51
lMEL	13.90	0.16
LSPF	13.05	0.85
SBPCA [26]	11.55	-0.01
SOUs [19]	15.29	0.11

6. CONTRIBUTIONS AND CONCLUSIONS

In this paper, we describe and analyze a number of techniques that when taken together performed best in the audio-only conditions of the 2013 TRECVID MED evaluation. Even after combination with video features, we observe decent gains, in particular for conditions where no (EK 0) or only

Table 4. TRECVID mAP results for pre-specified (top) and ad-hoc events (bottom), on MED13 evaluation data (left) and progress sub-set (right). We show audio and video features ("A-V"), video-only ("Visual"), non-ASR audio features ("Audio"), or ASR using delayed segmentation ("ASR"). "A-V" is a fusion of the "Visual", "ASR", and "Audio" systems, so the individual contributions can be compared.

mAP	EK100	EK10	EK0	EK100	EK10	EK0
A-V	30.6	12.6	3.7	36.3	21.2	10.1
Visual	26.4	11.2	2.4	28.4	16.3	5.2
ASR	7.8	2.0	1.8	5.7	2.6	3.1
Audio	12.6	4.7	0.3	16.1	9.1	0.2
A-V	33.4	12.8	4.7	36.0	22.5	10.6
Visual	28.1	11.6	3.0	28.7	18.4	5.5
ASR	9.8	1.8	2.1	5.2	2.0	2.1
Audio	13.6	4.8	0.4	16.8	8.6	0.3

few example (EK 10) videos are available to train a low-level retrieval model. Audio-based features and techniques are poised to gain importance, as "semantic" techniques become more relevant in the quest for compact representations and meaningful labels in multi-media data.

The present paper describes three techniques that distinguish our approach from its competitors, and evaluates their performance and relative contributions on top of a state-of-the-art baseline. The first is a transformation of lMEL features using sparse denoising auto-encoders, thereby applying deep learning principles to audio features without supervision. Second, we present that wide context LSPF extraction does not perform too well on its own, but it provides a complementary feature at the fusion stage. The third main contribution is a novel "delayed" segmentation technique, which optimizes the output of an ASR system for retrieval performance, rather than word error rate. This also relies on the LSP features, and significantly improved the detection of audio semantic concepts in consumer-grade audio.

In future work, we intend to further investigate the deep learning based features, and will attempt to perform semi-supervised training of semantic audio models by exploiting co-training with visual semantic features, to further improve the performance specifically for the few exemplar conditions. We expect that temporal modeling of sounds, or even a temporal model of entire events, should also lead to further gains.

7. ACKNOWLEDGMENTS

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding 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.

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

The authors would like to thank Shoou-I Yu and Zhen-Zhong "Danny" Lan for providing the fusion results for the audio-only and audio-visual systems presented in this paper.

8. REFERENCES

[1] Paul Over, Jon Fiscus, and Greg Sanders, "TRECVID 2013 – An introduction to the goals, tasks, data, evaluation mechanisms, and metrics," in *Proc. TRECVID*, Gaithersburg, MD; U.S.A., Nov. 2013, National Institute of Standards and Technology, http://www-nlpir.nist.gov/projects/tv2013/.

[2] Stephanie Strassel, Amanda Morris, Jonathan Fiscus, Christopher Caruso, Haejoong Lee, Paul Over, James Fiumara, Barbara Shaw, Brian Antonishek, and Martial Michel, "Creating HAVIC: Heterogeneous audio visual internet collection," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012, European Language Resources Association (ELRA).

[3] Geoffrey E. Hinton, Li Deng, Dong Yu, George E. Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[4] Susanne Burger, Qin Jin, Peter F. Schulam, and Florian Metze, "Noisemes: Manual annotation of environmental noise in audio streams," Tech. Rep. CMU-LTI-12-07, Carnegie Mellon University, Pittsburgh, PA; U.S.A., 2012.

[5] Qin Jin, Peter F. Schulam, Shourabh Rawat, Susanne Burger, Duo Ding, and Florian Metze, "Event-based video retrieval using audio," In *Proc. INTERSPEECH* [27].

[6] Benjamin Elizalde, Mirco Ravanelli, and Gerald Friedland, "Audio concept ranking for video event detection on user-generated content," in *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France, Aug. 2013, ISCA.

[7] Stephanie Pancoast, Murat Akbacak, and Michelle Sanchez, "Supervised acoustic concept extraction for

multimedia event detection," in *ACM Multimedia Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis (AMVA)*, Nara; Japan, Oct. 2012, ACM.

[8] Liangliang Cao, Shih-Fu Chang, Noel Codella, Courtenay Cotton, Dan Ellis, Leiguang Gong, Matthew Hill, Gang Hua, John Kender, Michele Merler, Yadong Mu, Apostol Natseve, and John R. Smith, "IBM Research and Columbia University TRECVID-2011 Multimedia Event Detection (MED) System," in *Proc. TRECVID*, Gaithersburg, MD; U.S.A., Nov. 2011, National Institute of Standards and Technology, http://www-nlpir.nist.gov/projects/tv2011/.

[9] Shourabh Rawat, Peter Schulam, Susanne Burger, Duo Ding, Yipei Wang, and Florian Metze, "Robust audio codebooks for large scale event detection in consumer videos," in *Proc. INTERSPEECH*, Lyon; France, Aug. 2013, ISCA.

[10] Stephanie Pancoast and Murat Akbacak, "Bag-of-audio-words approach for multimedia event classification," In *Proc. INTERSPEECH* [27].

[11] Xiaodan Zhuang, Stavros Tsakalidis, Shuang Wu, Pradeep Natarajan, Rohit Prasad, and Prem Natarajan, "Compact audio representation for event detection in consumer media," In *Proc. INTERSPEECH* [27].

[12] Stavros Tsakalidis, Xiaodan Zhuang, Roger Hsiao, Shuang Wu, Pradeep Natarajan, Rohit Prasad, and Prem Natarajan, "Robust event detection from spoken content in consumer domain videos," In *Proc. INTERSPEECH* [27].

[13] Vasant Manohar, Stavros Tsakalidis, Pradeep Natarajan, Rohit Prasad, and Prem Natarajan, "Audio-visual fusion using bayesian model combination for web video retrieval," in *Proceedings of the 19th ACM International Conference on Multimedia*, New York, NY, USA, 2011, MM '11, pp. 1537–1540, ACM.

[14] National Institute of Standards of Technology, "2013 TRECVID Multimedia Event Detection Track," http://www.nist.gov/itl/iad/mig/med13.cfm, Aug. 2013, Last acccessed: April 15, 2014.

[15] FFmpeg, "A complete, cross-platform solution to record, convert and stream audio and video," http://www.ffmpeg.org/, Sept. 2013.

[16] Hagen Soltau, Florian Metze, Christian Fügen, and Alex Waibel, "A One-pass Decoder based on Polymorphic Linguistic Context Assignment," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, Madonna di Campiglio, Italy, Dec. 2001, IEEE.

[17] Karen Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, 1972.

[18] Jonas Gehring, Yajie Miao, Florian Metze, and Alex Waibel, "Extracting deep bottleneck features using stacked auto-encoders," In *Proc. ICASSP* [28].

[19] Sourish Chaudhuri and Bhiksha Raj, "Unsupervised hierarchical structure induction for deeper semantic analysis of audio," In *Proc. ICASSP* [28], pp. 833–837.

[20] František Grézl and Petr Fousek, "Optimizing bottleneck features for LVCSR," in *Proc. ICASSP*, Las Vegas, NV; U.S.A., Apr. 2008, IEEE.

[21] Stanley S. Stevens, John Volkman, and Edwin B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[22] Jonas Gehring, Yajie Miao, Florian Metze, and Alex Waibel, "Extracting Deep Bottleneck Features Using Stacked Auto-Encoders," In *Proc. ICASSP* [28].

[23] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010, Oral Presentation.

[24] Geoffroy Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Tech. Rep., IRCAM, 2004, http://recherche.ircam.fr/anasyn/peeters/ARTICLES-/Peeters_2003_cuidadoaudiofeatures.pdf.

[25] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, New York, NY; USA, 2010, MM '10, pp. 1459–1462, ACM.

[26] Courtenay V. Cotton and Dan P.W. Ellis, "Subband autocorrelation features for video soundtrack classification," In *Proc. ICASSP* [28], pp. 8663–8666.

[27] *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, Portland, OR; U.S.A., Sept. 2012. ISCA.

[28] *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013*, Vancouver, BC; Canada, May 2013. IEEE.