Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor

Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller
Machine Intelligence & Signal Processing Group, MMK
Technische Universität München
80290 Munich, Germany
eyben@tum.de

ABSTRACT

We present recent developments in the openSMILE feature extraction toolkit. Version 2.0 now unites feature extraction paradigms from speech, music, and general sound events with basic video features for multi-modal processing. Descriptors from audio and video can be processed jointly in a single framework allowing for time synchronization of parameters, on-line incremental processing as well as off-line and batch processing, and the extraction of statistical functionals (feature summaries), such as moments, peaks, regression parameters, etc. Postprocessing of the features includes statistical classifiers such as support vector machine models or file export for popular toolkits such as Weka or HTK. Available low-level descriptors include popular speech, music and video features including Mel-frequency and similar cepstral and spectral coefficients, Chroma, CENS, auditory model based loudness, voice quality, local binary pattern, color, and optical flow histograms. Besides, voice activity detection, pitch tracking and face detection are supported. openSMILE is implemented in C++, using standard open source libraries for on-line audio and video input. It is fast, runs on Unix and Windows platforms, and has a modular, component based architecture which makes extensions via plug-ins easy. openSMILE 2.0 is distributed under a research license and can be downloaded from http://opensmile.sourceforge.net/.

Categories and Subject Descriptors

H.5 [**Information Systems Applications**]: Sound and Music Computing

General Terms

Design, Performance

Keywords

audio features, video features, multimodal fusion, real-time processing

MM'13, October 21–25, 2013, Barcelona, Spain.

ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502224.

1. INTRODUCTION

Feature extraction is an essential part of many audio and multimedia retrieval tasks, including automatic speech recognition, speaker characterization (by traits such as age and gender, or states such as emotion or sleepiness), music information retrieval (such as genre or music mood recognition), or video information retrieval (such as tagging of web videos or violence detection in movies). openSMILE aims at uniting features from these worlds, allowing researchers from either domain to profit from state of the art speech, music and video feature extraction. openSMILE provides a simple, scriptable console application where modular feature extraction components can be freely configured and connected via a text-based configuration file. Besides flexibility, sharing of such configuration files among researchers is believed to foster reproducibility of results. A strong focus is put on fully supporting real-time, incremental processing, allowing the use of openSMILE as front-end for interactive applications such as dialogue systems (cf. the SEMAINE project[1] for a dialogue system prototype utilizing openSMILE). At the time of this writing, openSMILE has been used as official baseline for the series of INTERSPEECH challenges in the area of Computational Paralinguistics [7, 8] where it has been shown to deliver state-of-the-art results approaching very diverse recognition tasks with a unified feature extraction scheme. Furthermore, it has been used by researchers (other than the authors) in more than 50 accepted research papers[2].

SMILE is an acronym for *Speech and Multimedia Interpretation by Large-space Extraction.* The term *large-space extraction* refers to openSMILE's capability of 'chunking' of the features in various ways so as to provide either frame-wise features, summarization by statistical functionals over parts of the input or the whole input, or any combination thereof. This kind of systematic feature (over-)generation enables explorative analysis in a very wide range of audio(-visual) recognition tasks. Even though openSMILE originates from the audio processing domain – as such, it has been featured in the 2010 ACM MM Open Source Software Competition [4] – it has recently been extended with basic video features, and, more importantly, its design is principally modality independent. For instance, also physiological measurements such as heart rate or EEG signals can be analyzed, e.g. by means of statistics and short-time spectral analysis. An easy to use plugin interface moreover provides the ability to extend openSMILE with one's own components, thus virtually being

[1] `http://www.semaine-project.eu/`
[2] According to Google Scholar citations, May 2013

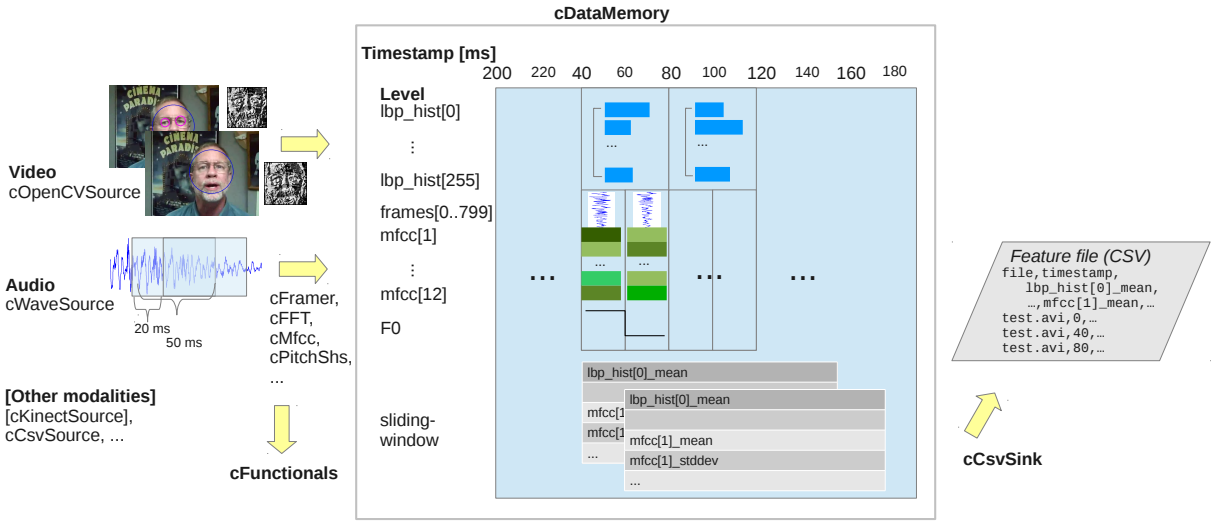

Figure 1: Exemplary audio-visual feature extraction process implemented using the openSMILE architecture. Audio and video frames are captured at 50/25 fps and stored along with timestamps in the data memory ring buffer. From the extracted low-level descriptors (LBP histogram, MFCCs 1–12 and pitch), the mean over 120 ms sliding windows is computed and written to a CSV file.

able to solve any feature extraction task and thereby using existing components as building blocks.

2. DESIGN AND FUNCTIONALITY

To meet our requirements, the following key design principles are implemented – a more detailed description can be found in the release documentation on the sourceforge project page[3]– : incremental ('on-line') processing from audio and video streams, as well as batch ('off-line') processing of audio, video and pre-extracted feature files; fast and lightweight algorithms implemented in C/C++ and tested against reference implementations; modular architecture allowing for arbitrary feature combination as well as integration of run-time plug-ins (e.g., for visualization or use in a dialogue system); scripting via configuration files to facilitate sharing and standardization of feature sets.

In the architecture of openSMILE, the *Data Memory* is the central link between all *Data Sources* (components which read data from external sources), *Data Processors* (components which copy and modify data), and *Data Sinks* (components which write data to external places such as files, or perform classification). Figure 1 visualizes an exemplary feature extraction process from audio and video. Audio, usually sampled at 16–44.1 kHz, is split into overlapping frames of 50 ms length at 20 ms frame rate (50 fps), and cepstral coefficients (MFCCs 1–12) and pitch (F0) are extracted. From the video, the local binary pattern (LBP) histogram is computed per video frame at 40 ms frame rate (25 fps) from faces detected from the frames converted to grey scale. The feature extraction process creates various *levels* of different size and frame rate in the data memory, as shown in detail in Figure 1. In particular, the framed audio is converted to 'row vectors' of size 16 kHz × 50 ms = 800 for subsequent processing, which are stored in the data memory for maximum flexibility: Any component can use the framed audio to apply its own processing steps. In contrast, the video source is implemented in a 'monolithic' fashion which does not involve storing the results of intermediate processing steps, such as conversion to grey scale, etc., for efficiency reasons. The frame-wise features such as MFCCs will subsequently be referred to as *low-level descriptors* (LLD).

In the example, statistical functionals (e.g., mean) are subsequently applied to the low-level descriptors over sliding windows of length 120 ms with 40 ms shift (in practice, one would usually use longer sliding windows). The data memory implements a first-in first-out (FIFO) ring-buffer architecture (hinted at by the timestamps displayed in Figure 1), where buffer sizes are determined automatically from the dependency graph of the components, by the amount of context that components need (e.g., computing the mean F0 over a sliding window of 120 ms as in the example requires at least 6 + 1 = 7 audio frames to be buffered at a frame step of 20 ms). In addition, buffers of unbounded size are supported in case that a component requests the full input (e.g., for cepstral mean normalization). Details on the ring-buffer architecture can be found in [4] and on the project webpage. In addition, recently, a *multi-loop processing mode* has been introduced where components are able to reset the input pointer after reading the full input. In audio signal processing, this is useful, e.g., for de-noising and de-reverberation by spectral subtraction and long-term cepstral normalization. Finally, the computed sliding window means are written to a comma-separated values (CSV) file with header, along with timestamps referring to the start of the sliding window.

Let us conclude the discussion of functionality by summarizing the LLDs, functionals, and export functionalities (sinks) implemented in openSMILE. The LLD currently implemented are listed in Table 1. The Mel-frequency features, Mel-Spectrum and Mel-Frequency Cepstral Coefficients (MFCC), as well as the Perceptual Linear Predictive Coefficients (PLP) are extracted in full accordance with the popular Hidden-Markov Toolkit (HTK) [12]. Extraction of visual descriptors is currently provided by integration of OpenCV routines. Face detection and normalization by eye

[3]http://opensmile.sourceforge.net/

alignment is supported prior to feature extraction. From the LLD contours, delta regression (Δ) coefficients [12] can be computed, and contours can be smoothed. The functionals in Table 2 can be applied to the LLD and Δ LLD contours to summarize time series in fixed length vectors. Export to popular toolkit formats such as HTK and Weka's Attribute Relation File Format (ARFF) [11] is supported besides 'dumping' the features in binary or text format.

Feature Group	Description
Waveform	Zero-Crossings, Extremes, DC
Loudness	Energy, intensity, *auditory model loudness*
FFT spectrum	Phase, magnitude (lin., dB, dBA)
ACF, Cepstrum	Autocorrelation and Cepstrum
Mel/Bark spectr.	Bands 0-N_{mel}
Semitone spectr.	FFT based and filter based
Cepstral	Cepstral features, e.g. MFCC, PLP-CC
Pitch	F_0 via Autocorrelation and sub-harmonic summation, *smoothed by Viterbi algorithm*
Voice Quality	HNR, Jitter, Shimmer, Voice Prob.
LPC	LPC coeff., reflect. coeff., residual Line spectral pairs (LSP)
Auditory	Auditory spectra, *psychoacoustic sharpness*
Formants	Centre frequencies and bandwidths
Spectral	Energy in N user-defined bands, roll-off points, centroid, entropy, flux, and rel. pos. of max./min., *harmonicity*
Tonal	CHROMA, CENS, CHROMA-based features
Color	*HSV colorspace histogram*
Texture	*Local binary pattern histogram*
Motion	*Optical flow histogram*

Table 1: openSMILE's low-level descriptors. *Italics* denote new components in version 2.0.

Category	Description
Extremes	Extreme values, positions, and ranges
Means	Arithmetic, quadratic, geometric
Moments	Std. dev., variance, kurtosis, skewness
Percentiles	Percentiles and percentile ranges
Regression	Linear and quad. approximation coefficients, regression err., and centroid
Peaks	Number of peaks, mean/stddev peak distance, mean/stddev peak amplitude, etc. (*more robust peak picking algorithm in 2.0*)
Segments	Number of segments based on delta thresholding *or various fixed thresholds*, mean/stddev. segment length, etc.
Sample values	Values of the contour at configurable relative positions
Times/durations	Up- and down-level times, rise/fall times, duration
Onsets	Number of onsets, relative position of first/last on-/offset
DCT	Coefficients of the Discrete Cosine Transformation (DCT)
LPC	*Autoregressive coefficients*
Zero-Crossings	Zero-crossing rate, Mean-crossing rate

Table 2: Functionals (statistical, polynomial regression, and transformations) available in openSMILE. *Italics* denote new components in version 2.0.

Category	Description
Feature files	Binary float matrix, CSV, HTK [12], Weka ARFF [11]
Classifiers / Regressors	LibSVM [1], *recurrent neural networks* [5], VAD

Table 3: Sinks available in openSMILE.

3. CASE STUDIES AND BENCHMARKS

Let us now briefly discuss a few case studies in order to illustrate the usage potential of openSMILE's features in a wide range of multimedia recognition tasks.

3.1 Paralinguistic Information Extraction

openSMILE has been used as the feature extractor for the official baselines of the 2009–2013 INTERSPEECH Challenges on paralinguistic information extraction, including emotion, age, gender, interest, alcohol intoxication, sleepiness, personality, voice pathology, and non-linguistic vocalization recognition [7, 8]. Building on the 2009–2012 Challenges' features and the lessons learnt, a unified feature set, the ComParE (Computational Paralinguistics Evaluation) set has been designed, comprising 6 373 features by brute-force combination of audio LLDs with functionals and subsequent removal of zero-information features, such as the arithmetic mean of delta coefficients which is usually near zero. Details on the feature set can be found in [8, 9]. The general applicability of the ComParE feature set in paralinguistic information extraction, not limited to speech, has recently been demonstrated: By regression on the ComParE features and subsequent feature selection, emotion can be determined from speech (GEMEP corpus), music (NTWICM corpus) and sound (ESD corpus) with state-of-the-art accuracy on speech and sound, and with so far best results on music (cf. [9] and Table 4). Details on the corpora and regression using Weka's [11] SVR can be found in [9]. The ComParE configuration file is delivered with the current release candidate.

3.2 Speaker Characterization in Web Videos

Let us now turn to multimedia feature extraction by considering speaker characterization in web videos as in [10]. An audio-visual feature extraction scheme similar to the one described in [10] using a proprietary implementation of video feature extraction can now be realized by using openSMILE exclusively. The underlying idea is close to the 'toy' example from Figure 1, yet extracting over 1.5 k LLD-functional combinations from the audio (INTERSPEECH 2010 set, IS10, delivered as configuration file with the openSMILE distribution) and using sliding window lengths of 4 s. The configuration file for synchronized audio and video feature extraction is also delivered with the current release candi-

date. In the result (cf. Table 4), 96.4 % and 70.0 % recall are obtained for gender and ethnicity, which is notable given the real-life nature of the task. Details on the classification procedure using Weka's [11] SVM can be found in [10].

3.3 Violence Detection

Finally, let us exemplify the use of openSMILE in multimedia retrieval by the 2012 MediaEval campaign's violence detection task in retail versions of Hollywood movies. openSMILE is used to extract a set of spectral and loudness descriptors (delivered with the source code distribution) over sliding windows of 2 s length with 0.5 s shift. Furthermore, mean and standard deviation of the optical flow and color (hue-saturation-value) histograms are added from the video frames. Again, classification is done using Weka's [11] SVM.

4. CONCLUSION AND OUTLOOK

The development of the openSMILE open source feature extractor, introduced for explorative feature generation from audio, stays very active. The high flexibility of the basic architecture has allowed for the integration of a variety of new functions compared to the 2010 release. Recent developments include multi-modal feature extraction by the addition of basic video features using integration of OpenCV; improved audio descriptors including, e.g., psychoacoustic features, and a multi-loop mode opening up a wide range of possibilities for more complex multi-pass feature extraction procedures; and context-sensitive recurrent neural networks as a powerful classifier and regressor. During the eNTERFACE 2012 workshop a Microsoft Kinect input component was developed. It supports depth and color image, as well as source audio after beamforming. The code is in an experimental stage, and will be made available soon.

Future developments will likely include a joint front-end for audio and video input such as ffmpeg, as well as the implementation of various on-line audio enhancement algorithms such as beamforming, non-negative matrix factorization (cf., e.g., [2]) and independent component analysis (cf., e.g., [6]). Moreover, a TCP/IP network interface will be available to facilitate transparent real-time interaction with distributed third-party systems. So far, openSMILE has been mostly taken up by the research community in the area of computational paralinguistics, where it has become a quasi-standard reference toolkit. We hope that many significant contributions, also from other communities, will follow.

Task	Modality/Features	Evaluation measure	
Emotion recognition [9]			Aro/Val
Speech	Audio/ComParE [8]	CC	.85/.50
Music	Audio/ComParE	CC	.67/.51
Sound	Audio/ComParE	CC	.59/.82
Speaker characterization in web videos [10]			
Gender	Video/LBP	UAR	78.2 %
	— +Audio/IS10 [7]	UAR	96.4 %
Ethnicity	Video/LBP	UAR	70.0 %
Violence detection in Hollywood movies [3]			
Violence	Audio/VSD_TUM [3]	MAP	48.4 %
Violence	Video/VSD_TUM [3]	MAP	32.0 %

Table 4: Task-based evaluation of openSMILE's feature extraction in 3 case studies (Sec. 3): Regression and classification of affect and speaker characteristics. CC: correlation coefficient, UAR: unweighted average recall, MAP: mean average precision.

5. ACKNOWLEDGMENT

The development of openSMILE has received funding from the European Commission (grant nos. 211486 and 289021, SEMAINE and ASC-Inclusion). The authors would like to thank Christian Kirst for helpful contributions.

6. REFERENCES

[1] C.-C. Chang and C.-J. Lin. *LibSVM: a library for support vector machines*, 2001. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] Z. Duan, G. J. Mysore, and P. Smaragdis. Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments. In *Proc. of Interspeech*, Portland, OR, USA, 2012.

[3] F. Eyben, F. Weninger, N. Lehment, G. Rigoll, and B. Schuller. Violent Scenes Detection with Large, Brute-forced Acoustic and Visual Feature Sets. In *Proceedings MediaEval 2012 Workshop*, Pisa, Italy, October 2012. 2 pages.

[4] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of ACM MM*, pages 1459–1462, Florence, Italy, October 2010. ACM.

[5] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[6] R. Maas, A. Schwarz, Y. Zheng, K. Reindl, S. Meier, A. Sehr, and W. Kellermann. A Two-Channel Acoustic Front-End for Robust Automatic Speech Recognition in Noisy and Reverberant Environments. In *Proc. of CHiME*, pages 41–46, 2011.

[7] B. Schuller. The Computational Paralinguistics Challenge. *IEEE Signal Processing Magazine*, 29(4):97–101, July 2012.

[8] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, et al. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proc. of INTERSPEECH*, Lyon, France, August 2013. ISCA. in press.

[9] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer. On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common. *Frontiers in Emotion Science*, 2013. DOI: 10.3389/fpsyg.2013.00292, in press.

[10] F. Weninger, C. Wagner, M. Wöllmer, B. Schuller, and L.-P. Morency. Speaker Trait Characterization in Web Videos: Uniting Speech, Language, and Facial Features. In *Proc. of ICASSP*, Vancouver, Canada, May 2013. IEEE. in press.

[11] I. H. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques, 2nd Edition.* Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK book (v3.4)*. Cambridge University Press, 2006.