xLiMe
FP7-ICT-2013-10
ICT-2013.4.1
Contract no.: 611346
www.xlime.eu

Deliverable D3.1.1

Early Prototype for Audio Annotation

Editor:	Nicu Sebe, UNITN
Author(s):	Dubravko Culibrk, UNITN; Nicu Sebe, UNITN
Deliverable Nature:	Prototype (P)
Dissemination Level:	Public (PU)
Contractual Delivery Date:	M12 – 31 October 2014
Actual Delivery Date:	M12 – 31 October 2014
Suggested Readers:	
Version:	1.0
Keywords:	audio annotation; early prototype; OpenSMILE

Disclaimer

Full Project Title:	xLiMe– crossLingual crossMedia knowledge extraction
Short Project Title:	xLiMe
Number and Title of Work Package:	WP3 Cross-lingual Multimedia Semantic Annotation
Document Title:	D3.1.1 - Early Prototype for Audio Annotation
Editor:	Nicu Sebe, UNITN
Work Package Leader:	Nicu Sebe, UNITN

Executive Summary

The main goal of the xLiMe project is to enable extraction of knowledge from different media channels and languages and relating this knowledge to cross-lingual, cross-media knowledge bases. The functional requirements for an early prototype of a system able to do this have been gathered and systemized in deliverable (D1.4.1) of xLiMe. An integral part of the system described there is the module for annotating audio in terms of emotions expressed by the speaker, which is the focus of task T3.1 of the project.

To meet the functional requirements for early audio annotation, as specified in D1.4.1, we need to develop tools to perform lightweight approximate annotation of audio streams in terms of emotional content. The goal is to extract low-level audio features from the information sources available on the brands and products of interest (e.g., IPTV streams from ZATTOO) and to calculate arousal and valence levels of the associated audio streams. These then need to be mapped onto a set of keywords with predetermined emotional interpretations (e.g., like, dislike).

Table of Contents

Abbreviations

MFCC	Mel-Frequency Cepstral Coefficients
IPTV	Internet Protocol television
HLS	HTTP Live Streaming

1 Introduction

This deliverable outlines the results of the scientific investigation undertaken in order to identify existing technologies related to the problem of emotion recognition in audio and use them to develop an early xLiMe prototype that will be able to do this within the scope of the xLiMe.

1.1 Background and Motivation

There are three use cases in xLiMe that will provide feedback end evaluation for the system to be developed. They will be run by the following partners:

- **ZATTOO [1]** is a pioneer of IPTV and leader in Switzerland and Germany, with additional presence in France, Spain, UK, Luxemburg, and Denmark. Zattoo earns income from ads, premium services, and B2B relationships. ZATTOO's proprietary technology assets include cloud-based recording which is currently configured to store over 200,000 hours (120 live TV channels are continuously recorded over the past 7 days and individual recordings for users in several countries) .
- **VICO Research & Consulting GmbH [2]** is concentrated on social media measurement and analysis, and the construction of social media monitoring systems as well as social media consulting. Their main customers are consumer goods manufacturers and marketing agencies. Clients amongst others are LG Electronics, Commerzbank, Symantec Europe, BMW, EnBW, Ferrero, Central, ENVIVAS, T-Systems, Mazda Europe, and Mindshare.
- **ECONDA GmbH [3]** focuses on web-analytics and recommendation solutions. For several years running, ECONDA is listed as one of the Top Five of web-analytics tools by independent experts. More than 1,000 satisfied e-business customers rely on ECONDA's web-analytics solutions. This includes customers such as retailers, textile specialists, manufacturers, brands, service providers, portals, publishers, price comparators, publishing houses, newspapers and NGOs. Since the Use Case of ECONDA builds on the other Use Cases and starts in Year2, details and requirements will be covered in D1.4.2.

The use cases have been carefully selected in order to demonstrate the advantages of the technology developed within the project. They will focus on two different applications of interest to different stakeholders:

- **Cross-media content enrichment and search**: Providing multimedia-content consumers with additional related content enhances the service provided by companies such as ZATTOO. The xLiMe project will develop applications, which will enable enrichment of the multimedia content of TV-channels watched by ZATTOO-users with related content, originating from other media sources (e.g., tweets, blog posts, YouTube videos, news articles, Wikipedia pages, etc.). The approach will be based on content, not on user behaviour.
- **Cross-media brand and topic monitoring**: The social media consulting process can be further enhanced by relating collections of social media documents, on a topic, to related TV-channels about the same topic. For instance to measure the coverage of topics in mainstream media which are trending in social media. From a business standpoint, brands are a topic of special interest for our use-case partners. The xLiMe project will provide tools to enable the annotation of mainstream multimedia streams with select advertisement presence and product placement data, which will then be used to establish the connection

between social and mainstream media. The annotation information is also useful to the multimedia content providers, as they will be able to analyse the product placement in the content. The annotation of the stream will be done by detecting logos, brands and ads in the multimedia stream and linking to the product shown in the ad. Initially, this will be done for a limited, predetermined number of logos, brands and ads.

The requirements for the early prototype of audio annotation component are mainly derived from the second (cross-media brand and topic monitoring) use case, where the component will be used to detect the sentiment of brand mentions, in the video stream.

While the focus of the early prototype is on brand-related data, the early prototype is designed to detect and extract as much sentiment data as possible from the audio stream, providing valuable input for both use cases.

The prototype harnesses the state-of-the-art technologies available and accessible to the consortium and builds upon them to provide a real-time performance sentiment annotation solution.

2 Sentiment Annotation from Audio

The pipeline of the xLiME early prototype for audio annotation is shown in Figure 1.

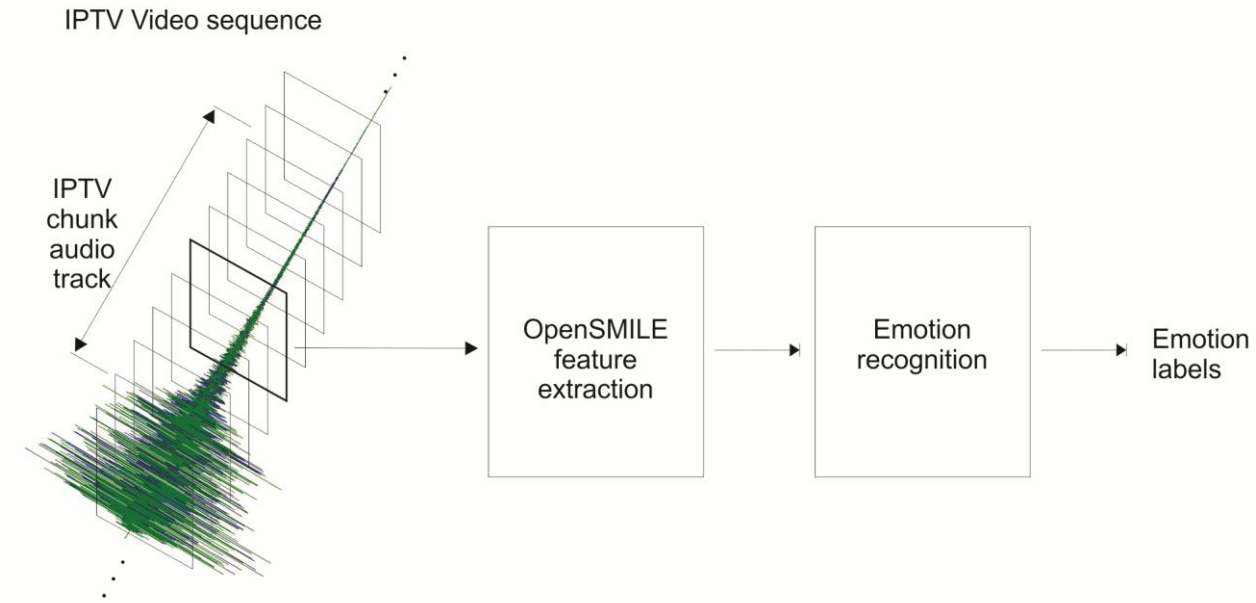

Figure 1: xLiMe audio annotation pipeline.

The system accesses the ZATTOO HLS stream, which provides the video data in 4 second chunks. The audio is extracted from the stream and fed to a basic audio feature extraction module [4].

The extracted features are then processed by the emotion recognition module which performs extraction of higher-level features and runs the emotion recognition classifiers. The processing in this module is done on 2 second frames. The classification for a frame is provided by binary classifiers trained to detect each emotion of interest. Each classifier predicts whether or not the current 2-second sliding frame is the characterized by the emotion in question or not. The results from the models are then compiled for the entire 4-second IPTV chunkand the label is obtained through majority voting.

The final outputs of the system are emotion labels for the audio chunk. The emotion classifier provides labels for “happiness/excitement”, “anger”, “sadness”, and “neutral”.

2.1.1 Feature Extraction

The basic feature extraction in the early prototype is performed using OpenSmile[4], which is an open source tool for audio feature extraction.

The openSMILE feature extraction tool enables us to extract large audio feature spaces in real time. SMILE is an acronym for Speech & Music Interpretation by Large-space Extraction, as it combines features from Music Information Retrieval and Speech Processing. The tool is written in C++ and is available as both a standalone command line executable as well as a dynamic library. Feature extractor components can be freely interconnected to create new and custom features, using a configuration file. The architecture is extendable and new components can be added using a simple binary plugin interface and a comprehensive API.

The initial set of features extracted with OpenSMILE consists of: Root Mean Square (RMS) energy, Mel-Frequency Cepstral Coefficients (MFCCs) (with 12 frames stacked), zero crossing rate, voice probability distribution, and fundamental frequency (F0) [4].

For each of these feature categories, further extended features are calculated in the final version of the prototype: maximum, minimum, range, maximum position, minimum position, mean, linear regression, standard deviation, skewness and kurtosis.

2.1.2 Training and Evaluation Data

To develop and validate the required early audio annotation prototype, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database.

IEMOCAP is an acted, multimodal and multi-speaker database, collected and published by the Speech Analysis and Interpretation Lab (SAIL) at University of Southern California. It contains approximately 12 hours of audio-visual data, which consists of 5 dyadic sessions where actors perform improvisations or scripted scenarios, specifically selected to elicit emotional expressions.

IEMOCAP database is annotated (using majority voting) by three annotators into categorical labels, such as anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation and dominance. Data for which no majority label exist is labelled is marked as ambiguous.

The database contains detailed motion capture information and the interactive setting to elicit authentic emotions. It represents one of the benchmark databases in the community for the study and modeling of multimodal and expressive human communication.

More information about the data can be found in [5].

2.1.3 Emotion Classification Experiments and Results

During the course of the development of the early prototype a number of experiments have been conducted.

Initial experiments were based on using simple classifiers and the basic features extracted with OpenSMILE to try to classify emotions. A decision table was created, which was able to correctly classifying 26% of the data. However, most of these instances were put into either the neutral or the ambiguous category.

Further experiments were conducted, where the ambiguous data was assigned a label provided from a randomly suggested judge. A decision stump, a decision table, and a J48 decision tree were created, correctly classifying at 29%, 29%, and 20%, respectively.

In the next set of experiments, the focus has been put on the classification of only the top 6 most common emotion labels: neutral, frustration, anger, sadness, happiness and excitement. A J48 tree and a Last Absolute Deviation (LAD) [8] tree were created, correctly classifying at 21% and 29% respectively.

In the subsequent research, the goal was set to try to match the performance achieved by Rozgić *et al.* [6]. They managed to get 49.1% to 66.3% accuracy per emotion, using acoustic features and

an RBF-SVM classifier on the same dataset, but focusing only on sadness, anger, neutral, and happiness and excitement - the last two merged into a single emotion of happiness.

An SVM classifier, using pre-defined features extracted with openSMILE, achieved comparatively low per-frame f-measures from 7.4% to 44.2% for the four emotions, and 30-35% for the weighted average f-measure.

In the next efforts, four separate models were trained, one for each emotion: sadness, anger, happiness/excitement and neutral.

Each model performs a binary classification of whether or not the current 2-second frame is the emotion in question or not. The f-measures for the best models for each emotion (using 10-fold leave-one-speaker-out cross validation on the IEMOCAP data) are shown in Table 1.

Table 1: Initial emotion classification results early prototype

		F-measure	
	Prior	**"Yes"**	**"No"**
Anger (1103)	49.07%	90.50%	71.96%
Happiness/Excitement (1636)	50.55%	83.68%	70.41%
Neutral (1708)	43.73%	85.21%	69.57%
Sadness (1084)	57.17%	89.43%	73.25%

The "Yes" and "No" columns represent the f-measures for correctly classifying as the emotion or rejecting an emotion in a given 2-second frame, respectively. The numbers in parentheses, after the emotion, represent the number of utterances with that label which are independent of each other, in the whole dataset. The prior probability of finding the emotion ("yes") in the training set is 50.23% for all models on the phrase-level. The priors in the table above are the probabilities of finding the emotion considering the number of frames in a phrase.

To be able to compare better with the baseline, in the later evaluation, we switched from 2-second-frame-based to utterance-based prediction. The results from the models applied to all 2-second frames within the utterance are compiled for the entire utterance using majority voting. The prior probabilities of each emotion in the dataset, under these conditions, are shown in Table 2.

Table 2: Per utterance prior probability of different emotions in IEMOCAP

	Prior
Anger (1103)	19.94%
Happiness/Excitement (1636)	29.58%
Neutral (1708)	30.88%
Sadness (1084)	19.60%

Next we introduced the additional features described in Section 2.1.2 and, once again, evaluated different classifiers. The best emotion detection accuracy was achieved by random tree emotion-specific classifiers (44.41% to 49.66%), which are comparable to the state-of-the-art as described in works of Rozgić *et al.* (2012)[6] and Lee *et al.* (2011)[5].

These final classifiers and result represent the state of the early audio annotation prototype at the time of writing of this report.

3 Conclusions

This deliverable describes the research conducted and technologies used to develop the xLiMe early audio annotation prototype. It also provides results of the initial evaluation of the prototype in terms of emotion detection performance.

We have successfully developed an audio annotation prototype relying on open source audio feature extraction technology and emotion recognition classifiers developed during the course of the project. The early prototype matches the performance of the state-of-the art in terms of detecting “anger”, “neutral”, “sadness” and “happiness/excitement” and is able to annotate the audio data contained in the ZATTOO IPTV streams with appropriate emotional labels.

Further system-level evaluation of the performance of this component will be conducted in the future and the prototype will be updated as necessary.

References

[1] http://corporate.zattoo.com

[2] http://www.vico-research.com

[3] http://www.econda.com

[4] Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor", In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013. doi:10.1145/2502081.2502224.

[5] Lee, C.-C., Mower, E., Busson, C., Lee, S., & Narayanan, S., “Emotion recognition using a hierarchical binary decision tree approach”,ScienceDirect, 53, pp. 1162-1171, July 2011.

[6] Rozgić, V., Ananthakrishnan, S., Saleem, S., Kumar, R., Vembu, A., & Prasad, R,“Emotion recognition using acoustic and lexical features”,Interspeech 2012, Portland, Oregon, September 2012.

[7] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.

[8] Breiman, Leo, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. Classification and regression trees. CRC press, 1984.

