SEMAINE

THE SENSITIVE AGENT PROJECT

D2b

Final face and voice feature extraction component with incremental, (near) real-time processing

Date: 24 September 2010

Dissemination level: Public

ICT project contract no.	211486
Project title	**SEMAINE** **Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression**
Contractual date of delivery	*30 June 2010*
Actual date of delivery	*24 September 2010*
Deliverable number	D2b
Deliverable title	Final face and voice feature extraction component with incremental, (near) real-time processing
Type	Demonstrator
Number of pages	13
WP contributing to the deliverable	WP 2
Responsible for task	Björn Schuller (schuller@tum.de)
Author(s)	Florian Eyben, Hatice Gunes, Maja Pantic, Marc Schroeder, Björn Schuller, Michel Valstar, Martin Wöllmer.
EC Project Officer	Philippe Gelin

Table of Contents

1 Executive Summary

Sensitive Artificial Listeners (SAL) are virtual dialogue partners who, despite their very limited verbal understanding, intend to engage the user in a conversation by paying attention to the user's emotions and non-verbal expressions. The SAL characters have their own emotionally defined personality, and attempt to drag the user towards their dominant emotion, through a combination of verbal and non-verbal expression.

This report is part of the series of reports describing the implementation of SAL in system SEMAINE-3.0. The software described, and the full set of reports, can be downloaded from http://semaine.opendfki.de/wiki/SEMAINE-3.0/.

This report describes the current state of the face and voice feature extraction integrated into the SEMAINE system 3.0, namely, incremental audio processing with multi-threading support, voice activity detection with speaker adaptation, incremental keyword-spotting and detection of non-linguistic audio events such as laughter and sighs with enhanced robustness, speaker adaptation on the feature level, audio recording and logging of features, face detection, action unit detection, and global head motion estimation.

2 Functionality of the components

This section describes the functionality of the components in the SAL system. The possibilities to configure and reuse the components as parts of a research toolbox will be published as deliverable D7e in December 2010.

2.1 Real-time audio feature extraction architecture

The audio input module uses the TUM openSMILE toolkit for real-time audio processing and feature extraction as well as classification. It provides a flexible architecture in which the whole processing chain can be easily configured via a set of configuration files. The architecture of openSMILE is presented in [2], and the technical details are described in the openSMILE book, which can be downloaded from http://www.openaudio.eu.

The openSMILE module extracts features for the voice activity detector (Section 2.2), prosodic analysis (Section 2.3), the keyword spotter and non-linguistic vocalisation detector (Section 2.4 and 2.5), and the acoustic emotion recognition module (D3b report).

2.2 Voice Activity

The voice activity detection uses a Long-Short-Term Memory (LSTM) Recurrent Neural Network (RNN) [5] trained on the TIMIT corpus on a speech non-speech discrimination task. The network has two hidden layers, the first with 140 memory cells (one cell per block) and the second with 10 memory cells. Different noise types have been added to the training data. This enables noise robust user speech detection. The detector, however, tends to skip smaller turns and also triggers on agent speech coming from the system speakers.

To have a more accurate, user adapted speech detection, single Gaussian Models (GMs) are trained and updated incrementally during run-time for the user's voice, the agent voice, and the background noise. The output of the LSTM speech detector is used as a training target for the GM model updates. For updating the agent model and avoiding misclassification of the agent's own voice as user speech until the model is sufficiently trained, a callback message, indicating the start and stop of the agent's voice output, is read from the topic *semaine.callback.output.Animation*.

After 6 seconds of user speech the user speech model is used in addition to the LSTM output (both are or'ed). This enables detection of shorter turns and decreases the latency, since the LSTM detector has shown some latency due to its memory capabilities.

The result of the voice activity detector is sent as a speakingStatus event (start/stop) message to the topic *semaine.data.state.user.emma.nonverbal.voice*.

2.3 Prosody

The SEMAINE AudioFeatureExtractor component sends pitch contour, energy, and loudness to the other SEMAINE components. Moreover, a per pseudo-syllable estimate of pitch direction (rise, fall, rise-fall) is computed.

The pitch extraction algorithm is based on sub-harmonic summation (SHS). In order to reduce delay, no Viterbi-style post-processing is performed. A simple 3 frame smoothing strategy is applied instead.

Root-Mean-Square (RMS) and Log energy are extracted and sent to the system. Also a psychoacoustic intensity and loudness measure is computed from the audio signal using a narrow-band approximation (without spectral weighting).

The detection of pseudo-syllables is based on a probability of voicing measure inferred from the SHS pitch detection algorithm.

The prosodic features are sent as feature vectors at a constant frame rate of 100 frames per second to the system topic *analysis.features.voice*.

2.4 Incremental Keyword Spotting

The SEMAINE 3.0 keyword spotter is able to detect a set of 173 keywords which are relevant for the dialogue management and for linguistic emotion recognition [1]. For each keyword, the system outputs the exact timing as well as a confidence between 0.0 and 1.0. As system responses have to be prepared already before the user has finished speaking, the keyword spotter operates incrementally, meaning that the current best guess of the keywords contained in the utterance spoken so far is output and updated at a constant rate (the default rate is 600ms).

The keyword spotting module uses a feature set consisting of Mel-Frequency Cepstral Coefficients (MFCC) 0 to 12, together with first and second order regression coefficients, which are generated by the openSMILE feature extractor [2]. On-line cepstral mean normalization is applied to the MFCC features in order to remove the effects of stationary noise and varying channel characteristics. The applied speech decoder uses the Julius library[1] which supports two-pass decoding. Thus, it processes the speech feature vector sequence in forward and in backward direction and can output a refined final hypothesis once the complete speech turn is available (i.e. once a silence period is detected).

The acoustic models used by the keyword spotter were trained using all utterances from both, user and operator as recorded in the SEMAINE database (recordings 1 to 19), the SAL corpus, the COSINE corpus, and the non-linguistic vocalisations contained in the DFKI-London Dialogue Corpus. Based on these speech corpora, tied-state cross-word triphone models (built from a set of 39 monophones) were trained. All phoneme Hidden Markov Models (HMMs) consist of three states with 16 Gaussian mixtures each. Unlike the SEMAINE 2.0 keyword spotting system, which exclusively relied on acoustic modelling for vocabulary independent detection of keywords, the keyword detector used for the SEMAINE 3.0 system uses a trigram language model trained on spontaneous speech (as contained in the SEMAINE database, the SAL corpus, and the COSINE corpus).

A further advancement is the novel multi-stream technique used for the SEMAINE 3.0 keyword spotter. In contrast to the single-stream HMMs applied in the SEMAINE 2.0 system, the multi-stream architecture not only processes MFCC feature vectors, but also models a second stream which contains the maximum likelihood phoneme estimate generated by an on-line Long Short-Term Memory (LSTM) phoneme predictor integrated into openSMILE (see [3], for example). The

1 http://julius.sourceforge.jp/en_index.php

underlying LSTM network consists of 128 memory blocks that enable long-range context modeling for enhanced framewise phoneme prediction [4].

2.5 Automatic Detection of Non-Linguistic Vocalisations from Audio Cues

The acoustic non-linguistic vocalisation detector as contained in the SEMAINE 3.0 system supports three different non-linguistic vocalisations: 'breathing', 'laughing', and 'sighing'. Training material for 'coughing' was too sparse to be included in the model set. The non-linguistic vocalisation detector uses the same speech decoder framework as the keyword spotter (see D2b), meaning that non-linguistic vocalisations are integrated in the acoustic and language model used by the keyword spotter. For model training, the SEAMINE database, the SAL corpus, the COSINE corpus, and the DFKI-London Dialogue Corpus were used. Unlike the phoneme models, the HMMs used for modeling non-linguistic vocalisations consist of nine states.

In order to exploit the multi-stream LSTM-HMM technique (see D2b) also for the detection of non-linguistic vocalisations, the LSTM predictor has additional output nodes for 'breathing', 'laughing', and 'sighing'. Thus, the generation of framewise LSTM predictions (as detailed in [1], [3], and [4], for example) is not limited to phoneme classes but also supports non-linguistic vocalisations.

2.6 The VideoFeatureExtractor component

This component provides functionality for all tasks that need to work on the acquired image data itself. Where possible, tasks have been split into the generation of a low-dimensional signal from this image data, which can then be further analysed in another module. Good examples of this are the detection of nods and shakes and of head tilts, where information about the head motion and head pose are computed in the VideoFeatureExtractor module and sent to a specific topic so that the nodshakeAnalyser and headPoseAnalyser components can detect these gestures.

The VideoFeatureExtractor component sends four signals to the framework: the detected face location, 2D head motion estimation, head pose estimation, and appearance-based Action Unit (AU) detection. The face location and 2D head motion estimation have been explained in detail in SEMAINE-2.0.

The head pose estimation uses the face detection result combined with eye detection within the detected face region to estimate the 3-Dimensional head position relative to the camera as well as the head roll. The location of the detected eyes can be used to estimate the head roll α as follows:

$$\alpha = \arctan\left(\frac{y_l - y_r}{x_l - x_r}\right)$$

where x_l and x_r are the horizontal position of respectively the left and right eye, and y_l and y_r are their vertical positions. The head roll is send to the topic “semaine.data.analysis.features.video.headpose”, as a feature message.

The appearance-based Action Unit detection uses Uniform Local Binary Pattern histograms, that are computed from each cell of a 10 x 8 grid placed over the detected face. The resulting feature vector is given to a series Support Vector Machine (SVM) classifiers, one for every trained AU. Currently the system has trained SVMs for the AUs AU1 (raised inner eye-brow), AU2 (raised outer eye-brow), AU4 (lowered eye-brows), AU12 (smile), and AU25 (lips parted). [5]

2.7 Fusion of Non-verbal cues (NonverbalFusion component)

The NonverbalFusion component merges non-verbal cues from the audio and the video modality. A simple late fusion approach is implemented, where the events are forwarded by the fusion component if a certain confidence threshold is exceeded. For laughter with a low acoustic confidence (< 0.7), AU12 (smile) must have been detected within a 1 second time window in order for the non-verbal fusion component to not discard the event. Also, if sigh (confidence < 0.7) and AU12 are detected within a 1 second window, both events are discarded. Due to the buffering, the component adds a delay of max. 1 second to the non-verbal events with low confidence.

3 Quality assessment

This section describes an assessment of the quality on the technical and component level. An assessment of the psychological quality of interactions with the overall system will be published separately, as deliverable report D6d, in December 2010.

3.1 Incremental Keyword Spotting

In order to assess improvements over SEMAINE 2.0, both, the keyword spotter integrated into the SEMAINE 3.0 system and the SEMAINE 2.0 keyword detector were evaluated with respect to true positive and false positive rates, using all 173 SEMAINE keywords. Since the recordings 1 to 10 of the SEMAINE database have already been used for training the acoustic models used by SEMAINE 2.0, we tested keyword spotting performance on the recordings 11 to 19 of the SEMAINE database (only the utterances spoken by the user). The performance obtained for the SEMAINE 2.0 system, different variants of the SEMAINE 3.0 keyword spotter, as well as for a commercial dictation software (Vocon 3200, using standard British English acoustic and language models containing all keywords) can be seen in Figure 1.

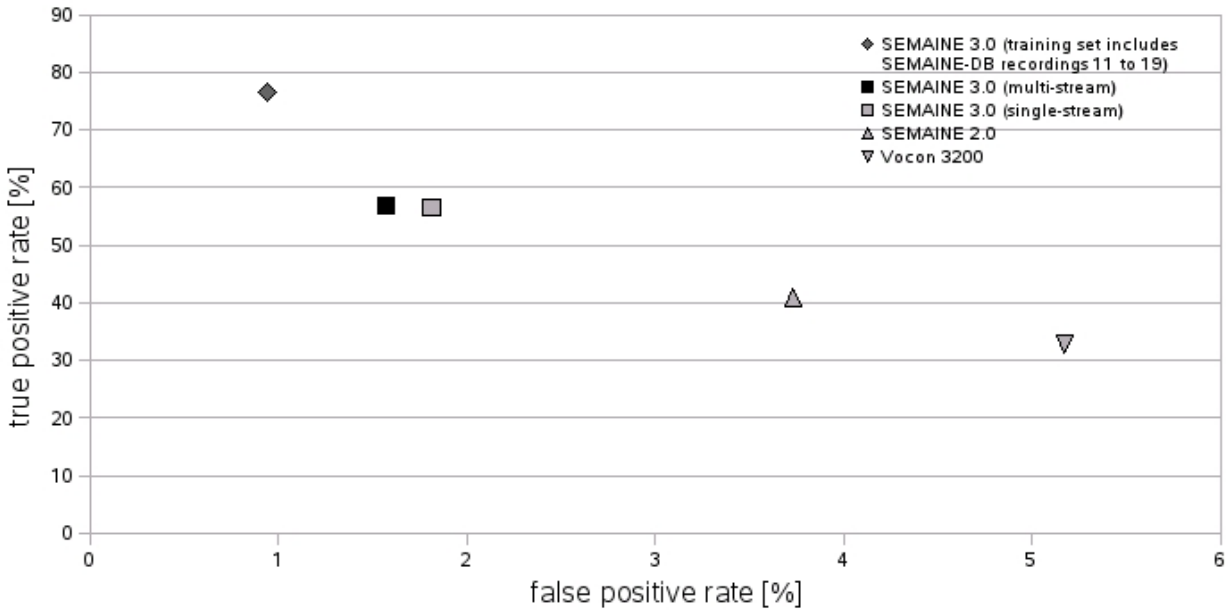

Figure 1: Keyword spotting performance of different variants of the SEMAINE 3.0 system, the SEMAINE 2.0 system, and a commercial keyword spotter based on the Vocon 3200 ASR engine.

When excluding the SEMAINE-DB recordings 11 to 19 (i.e. the test set) from the training set, the single-stream SEMAINE 3.0 keyword spotter achieves a true positive rate of 56.5% at a false positive rate of 1.81%. Such Receiver Operating Characteristic (ROC) operating points are typical for spontaneous, emotionally coloured speech [3] and keyword vocabularies containing very short words such as “I”, “up”, “to” etc. When using the multi-stream LSTM-HMM techinque, the false positive rate can be reduced to 1.57%.

Since the final acoustic and language models used for the SEMAINE 3.0 system also include the SEMAINE-DB recordings 11 to 19, a keyword spotter using these models was also tested for reference. This led to a true positive rate of 76.6.% and a false positive rate of 0.94%. Note, however, that this evaluation cannot be seen as a realisitc performance assesment since the test set was used to train the final SEMAINE 3.0 models (i.e. training and test set are not disjunctive in this case).

Both, the SEMAINE 2.0 keyword spotter and the keyword spotter based on a commercial dictation software (Vocon 3200) perform significantly worse than the SEMAINE 3.0 system.

3.2 Automatic Detection of Non-Linguistic Vocalisations from Audio Cues

The SEMAINE 3.0 non-linguistic vocalisation detector was evaluated with respect to true positive and false positive rates in order to assess improvements over SEMAINE 2.0. Since the recordings 1 to 10 of the SEMAINE database have already been used for training non-linguistic vocalisation models used by SEMAINE 2.0, we tested the performance on the recordings 11 to 19 of the SEMAINE database (only the utterances spoken by the user). The performance obtained for the SEMAINE 2.0 system and for different variants of the SEMAINE 3.0 system can be seen in Figure 1.

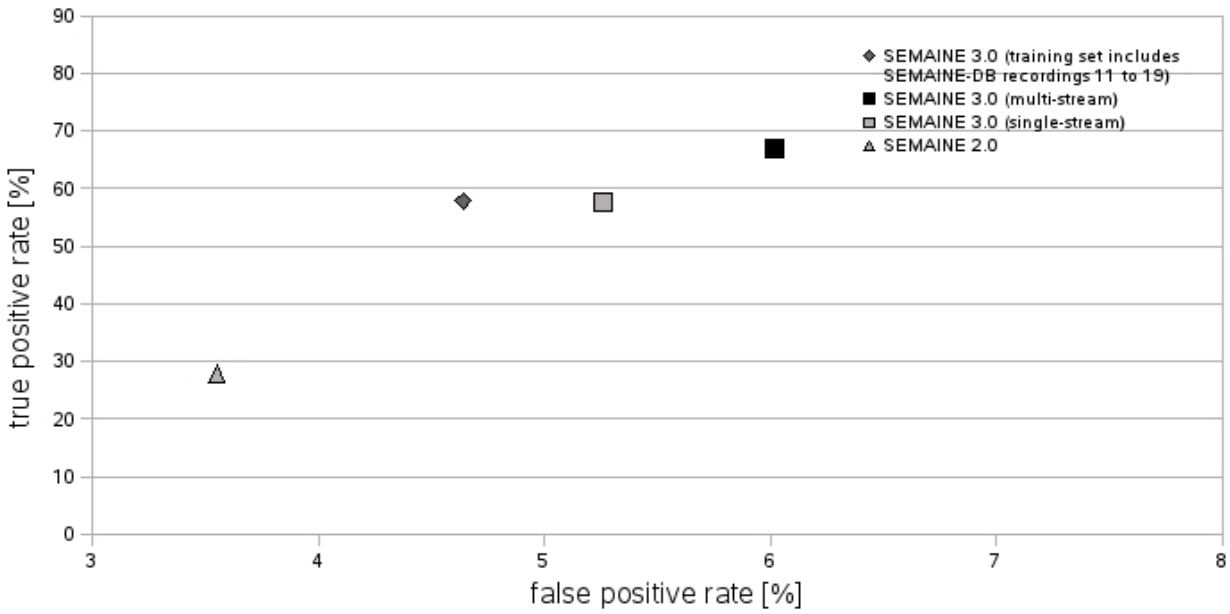

Figure 1: Non-linguistic vocalisation detection performance of different variants of the SEMAINE 3.0 system and the SEMAINE 2.0 system.

The preliminary non-linguistic vocalisation detector as contained in the SEMAINE 2.0 system leads to a rather low true positive rate (27.9%). The improvements made during the development of the SEMAINE 3.0 speech decoder lead to a higher true positive rate of 57.6%, at the expense of a higher false positive rate. A further improvement of the true positive rate can be obtained when using the multistream architecture applying Long Short-Term Memory modeling (66.9%).

Including SEMAINE-DB recordings 11 to 19 in the training set leads to a reduction of the false positive rate to 4.64%. However, as mentioned in D2b, this scenario is rather unsuited for a realisitc

performance assessment since training and test set are not disjunctive when testing the final SEMAINE 3.0 models on a fraction of the SEMAINE database.

3.3 Evaluation of the VideoFeatureExtractor component

The features extracted in this component all serve as input to analysers that perform the actual recognition of head- and face gestures. It is hard to evaluate the performance of this component independently of the analysers. Besides, the evaluation of the relevant analysers serve as sufficient evaluation of the quality of the extracted features. We therefore refer to the relevant sections in deliverable D3c for this.

4 License and availability

- Voice feature extraction is available under the terms of the GPL within the SEMAINE system (included in SEMAINE download package; linux and windows versions available); the voice feature extraction module is also available to the research community as an open-source standalone library and command-line tool, called openSMILE [2].
- Keyword spotting is based on the open-source Julius engine, which is available as a third-party download under a BSD-style license.
- Facial feature extraction modules are all part of the VideoFeatureExtractor module, which is available as binary component within the SEMAINE system download package under the terms of a binary-only research licence.

References

[1] Martin Wöllmer, Björn Schuller, Florian Eyben, Gerhard Rigoll: "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening", in IEEE Journal of Selected Topics in Signal Processing (J-STSP), Special Issue on Speech Processing for Natural Interaction with Intelligent Environments, IEEE, vol. 4, no. 5, 2010.

[2] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", in Proc. of ACM Multimedia, ACM, Firenze, Italy, 2010.

[3] Martin Wöllmer, Florian Eyben, Alex Graves, Björn Schuller, Gerhard Rigoll: "Bidirectional LSTM Networks for Context-Sensitive Keyword Detection in a Cognitive Virtual Agent Framework", in Cognitive Computation, Special Issue on Non-Linear and Non-Conventional Speech Processing, Springer, vol. 2, no. 3, pp. 180-190, 2010.

[4] Martin Wöllmer, Florian Eyben, Björn Schuller, Gerhard Rigoll: "Recognition of Spontaneous Conversational Speech using Long Short-Term Memory Phoneme Predictions", in Proc. of Interspeech 2010, ISCA, Makuhari, Japan, 2010.

[5] Bihan Jiang, "A comparison and investigation of dense temporal appearance descriptors for AU analysis", Master of Science Thesis submitted for the MSc degree in Computing Science of Imperial College London, September 2010.