SEMAINE

THE SENSITIVE AGENT PROJECT

D3a

Human conversational signals analyser

Date: 22 December 2009

Dissemination level: Public

ICT project contract no.	211486
Project title	**SEMAINE** **Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression**
Contractual date of delivery	*31 December 2009*
Actual date of delivery	*22 December 2009*
Deliverable number	D3a
Deliverable title	Human conversational signals analyser
Type	Demonstrator
Number of pages	9
WP contributing to the deliverable	WP 3
Responsible for task	Maja Pantic (m.pantic@imperial.co.uk)
Author(s)	Florian Eyben, Hatice Gunes, Dirk Heylen, Maja Pantic, Björn Schuller, 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.

The SEMAINE system 2.0 is the first public demonstrator of the fully operational autonomous SAL system based on audiovisual analysis and synthesis. The present report is part of a group of reports describing various aspects of the SEMAINE system 2.0. The full list of reports is available from http://semaine.opendfki.de/wiki/SEMAINE-2.0.

This report describes the current state of the human conversational signals analysers integrated into the SEMAINE system 2.0, namely, the automatic detection of head nods and shakes from visual cues, detection of interest from audio cues, detection of non-linguistic vocalisations from audio cues, detection of gender from audio cues, and prediction of the end of a user turn.

2 System description

2.1 Automatic Detection of Head Nods and Shakes from Visual Cues

Automatic detection of head nods and shakes from visual cues requires the existence of annotated nod and shake videos. To this aim, training data was obtained by visually inspecting SAL and SEMAINE databases and manually cutting 100 head nod and 100 head shake clips of variable length.

Automatic detection of head nods and shakes has then been attempted based on 2D global head motion estimation (explained in WP2). Based on the characteristics of the global head motion in every frame, geometric features namely, magnitude and direction (angle) were calculated. The extracted feature vectors were then fed into a Hidden Markov Model (HMM) for training a nod HMM and a shake HMM. The purpose of the two HMMs built in the previous step is to output the likelihood of a sequence being a nod or a shake. However, there exist other (somewhat noisy) head movements, that are neither a nod nor a shake, and that do not carry relevant information for automatic detection of nods/shakes. Such head movements need to be distinguished from the actual head nods/shakes. There is no straightforward solution to this problem. Our first solution consists of thresholding the magnitude of the head motion. Our second solution consists of building an 'other HMM' that is be able to recognize any movement but nods/shakes. The likelihood provided by this HMM is then used to make a decision on the nature of the analysed head movement. The final solution to the problem of differentiating nods/shakes from other head movement sequences was obtained by statically analysing the likelihoods outputted by the nod HMM and the shake HMM. Combining the results in this way proved more reliable than using a single classifier alone.

The head nod and shake detector has been integrated in the nodshakeAnalyser module of the SEMAINE framework. Information about the head nod and head shake is sent to the data.state.user-.emma topic as an EMMA message with a frequency of 0.6 seconds (i.e., every 30 frames). The EMMA message is only sent if a face has been detected previously.

2.2 Automatic Detection of Interest from Audio Cues

As for detection of head nods and shakes, detection of interest from audio requires annotated data. Here the TUM Audio-Visual Interest Corpus (AVIC, Schuller et al., 2009) was used. For this corpus, level of interest (LOI) labels exist on a chunk level (parts of turns or short turns) in three discrete classes, namely LOI 0, 1, and 2, which correspond to boredom and indifference, slight interest, and high interest. Preliminary analysis was conducted on the interest annotations in SEMAINE data.

A set of prosodic and spectral features (functionals of low-level descriptors) is used. The classifier of choice is Support Vector Machines (SVM) with a polynomial kernel function of degree 1. These have been shown as suitable for the problem in numerous publications (e.g. Schuller and Rigoll, 2009; Eyben et al., 2009).

Since some components in the SEMAINE system require a continuous level of interest value, the centroid of the class confidences (the SVM probability estimates) is used. This rather unconventional approach has in practice shown better performance than a regression approach, which usually would be used for continuous outputs. A regression model in the case of the AVIC database, however, can be trained only with the average of finite discrete LOI annotations, thus leaving gaps between discrete steps. This seems to heavily degrade performance because the model is not able to interpolate well enough. Training regression models on the new SEMAINE data is part of current work.

The interest detector is currently part of the SEMAINE component TumFeatureExtractor. The level of interest and the confidences for the three discrete levels is sent in an Emma XML message containing EmotionML to the topic state.user.emma.

2.3 Automatic Detection of Non-Linguistic Vocalisations from Audio Cues

For the automatic recognition of non-linguistic vocalisations we use HMMs trained on the SAL, SEMAINE, AMIDA, and AVIC database. Currently, the following non-linguistic vocalisations are supported: laughing, breathing, coughing, hesitating, and sighing. In conformance with experiments and optimisations in (Schuller et al., 2008), we apply left-to-right HMMs with nine states for every non-linguistic vocalisation.

Even though the HMM topology used for the detection of non-linguistic vocalisations differs from the model topology applied for phoneme and keyword decoding, respectively, the non-linguistic vocalisations decoder is integrated into the keyword spotter component and therefore uses the Julius library for decoding. Thus, the keyword spotting module outputs unified EMMA messages containing not only keywords but also non-linguistic vocalisations, together with their start times and confidences. As mentioned in D2, these messages are sent to the topic state.user.emma.

2.4 Automatic Gender Recognition from Audio Cues

Gender models based on acoustic features were trained on the following emotional speech corpora: the SEMAINE database, the SAL corpus, the Vera-Am-Mittag (VAM) corpus, the TUM Audiovisual Interest corpus, the TUM Airplane Behaviour Corpus (ABC), the Berlin Speech Emotion Database, and the ENTERFACE'05 database.

In contrast to existing gender recognition systems which build frame level models (usually Gaussian Mixture Models (GMM), or Neural Networks), we decided to take an alternative approach of turn level modelling with Support Vector Machines. This allows for re-use of the feature set and the techniques used for classification of interest and emotional state. No additional effort is required on the feature extraction side. A reduced feature set of 30 features was found by automatic feature selection and transformation (Correlation based feature subset selection and principal component analysis) followed by manual inspection and analysis. The set includes the 3% percentile of the fundamental frequency envelope, the 97% percentile of the voicing probability, means and moments of higher order MFCC coefficients (11-16), and energy distribution in the 50-150Hz band. With the described system an accuracy of 92.5% can be achieved with Support Vector Machines and a polynomial kernel function of degree 2 for the emotionally coloured speech. The SVM complexity factor was 0.8. In contrast to detecting valence and arousal from facial points (see SEMAINE report D3b), for audio RBF kernel functions resulted in performance slightly lower than the polynomial kernel function with more computational effort. Thus, we prefer the polynomial kernel. The result is below state of the art systems which have error rates of only a few percent or less owed to the fact that these are usually not evaluated on emotionally coloured speech which is a considerably more challenging (Schuller et al., 2009b); however, this is sufficient for our system, and it comes without additional computation costs and low implementation effort. In the SEMAINE system the final gender estimate is refined over the first 7 turns by majority voting.

The gender recogniser is currently part of the SEMAINE component TumFeatureExtractor. The refined gender result is sent to the topic state.user.emma in an Emma XML message containing SemaineML.

2.5 Detection of pitch variations

Based on the low level features *probability of voicing* and *fundamental frequency* pseudo syllable units are identified. These units correspond to continuous voiced sections. Thereby the minimum unvoiced time between two voiced segments must be 20 ms, otherwise the two segments are treated as one.

For each pseudo syllable segment pitch direction statistics are computed using the differences between a long term average and a short term average moving window. Furthermore, the short term average at the beginning and end of each segment is compared. Based on this analysis, using a relative minimum variation threshold, a decision for five classes of pitch variation can be performed per syllable. These classes are: flat, rising, falling, rise-fall, fall-rise. Most important for the dialogue are the classes rise and fall. Since flat is usually the most frequent case, messages are only sent for the last four cases. The message for the pitch direction event is an Emma XML message containing SemaineML. The message is sent to the topic state.user.emma every time a pitch variation event is detected.

From the pseudo syllable segments a pseudo syllable rate is computed on a per turn basis. However, this rate is not yet used in the system, and thus, not sent, yet.

3 License and availability

The automatic head nod and shake detector from visual cues is available as freeware as a binary executable functioning as part of the overall SEMAINE system 2.0. It is also available separately, as an installable executable together with all other components that rely on visual information processing and analysis.

Detection of interest and gender recognition is available under the terms of the GPL within the SEMAINE system (included in SEMAINE download package; linux and windows versions available); the classification uses the open-source third-party library LibSVM for support vector classification, which is distributed under a BSD-style license.

Detection of non-linguistic vocalisations uses the open-source Julius engine, which is available as a third-party download under a BSD-style license.

References

(Eyben et al., 2009) F. Eyben, M. Wöllmer, and B. Schuller, “openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit”, Proc. of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009), Amsterdam, The Netherlands, IEEE, pages 576-581, 2009.

(Schuller et al., 2008) B. Schuller, F. Eyben, and G. Rigoll, “Static and Dynamic Modelling for the Recognition of Non-Verbal Vocalisations in Conversational Speech”, Proc. of 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-based Systems (PIT 2008), Kloster Irsee, Germany*,* vol. LNCS 5078, pp. 99–110, Springer, 2008.

(Schuller and Rigoll, 2009) B. Schuller, and G. Rigoll, “Recognising Interest in Conversational Speech - Comparing Bag of Frames and Supra-segmental Features”, Proc. Interspeech (2009), Brighton, UK, ISCA, pp. 1999-2002, 2009.

(Schuller et al., 2009) B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, “Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application”, in: Image and Vision Computing Journal (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior, vol. 27, no. 12, pp. 1760-1774, 2009.

(Schuller et al., 2009b) B. Schuller, M. Wöllmer, F. Eyben, and G. Rigoll: “Retrieval of Paralinguistic Information in Broadcasts”, in Multimedia Information Extraction, Mark Maybury (ed.), MIT Press, Cambridge, Massachusetts, 2009.