Detection of Cognitive States and Their Correlation to Speech Recognition Performance in Speech-to-Speech Machine Translation Systems

Hayakawa Akira, Fasih Haider, Loredana Cerrato, Nick Campbell, Saturnino Luz

School of Computer Science and Statistics, Trinity College Dublin, Ireland
{campbeak, haiderf, cerratol}@scss.tcd.ie, nick@tcd.ie, luzs@cs.tcd.ie

Abstract
An analysis of possible associations between speech recognition performance and three cognitive states that arise in dialogues mediated by a speech-to-speech machine translation system is reported. This analysis is based on a new corpus of interlingual interactions in a map task which includes precisely synchronised speech, video, and physiological data streams (blood-volume pulse, skin conductance, electroencephalogram, and eye movements). While no evidence is found that cognitive states occurring prior to utterances sent to the speech recogniser affect the speech recognition performance, the onset of cognitive states – especially frustration – is found to be clearly associated with speech recognition performance. Given this association, methods for automatic detection of these cognitive states were explored by using features of the two physiological signals, features of the speech signal, and combinations of speech and physiological features. Combined biosignals yields detection performance well above the baseline (71% accuracy) when the time window is restricted to the perceived duration of the state. Extending the window to the end of the utterance following the cognitive state yields poor detection on biosignals alone, but improves considerably when features of the speech signal are added, thus showing the potential usefulness of speech features as a biosignal.

Index Terms: speech recognition, human-computer interaction, cognitive states

1. Introduction

Over the past 15 years, computer and speech scientists have explored various methodologies to automate the process of emotion, affective and cognitive state recognition. Past research has mostly focused on emotion recognition from one single sensorial source, or modality: mainly the face [1]. Given the fact that emotion, affective and cognitive states of a user can influence the unfolding of the interaction with the system, increasing effort has been spent to test methods for recognition and detection of different affective and cognitive states in human-machine communication.

Detecting the cognitive reactions of a user could be a step forward in the process of designing proactive systems capable of adapting to the user’s needs [2]. While it is true that the face is the main display of a person’s affective and cognitive state, other sources such as body movements and gestures have been shown to increase the recognition accuracy [3], [4], [5], [6], [7], [8] and achieve better results in the prediction of user’s affective and cognitive reactions. Similarly, features of the speech signal itself have been employed in inferring what has been loosely termed “emotion” or “affect” in the literature [9].

While great progress has been made in recent years on detecting such cognitive states from speech and other modalities on a number of speech datasets, the data used have mostly come from acted speech collected in non-interactive settings [9]. Studies involving the dynamics of cognitive states in interactive systems, specially systems where the interaction is mediated by automatic speech recognition (ASR) have been far less common. In speech-to-speech machine translation (S2S MT), the limitations of the technologies involved both ensure the elicitation of certain linguistic and cognitive reactions (in response to ASR and MT errors) and require careful design to address communication issues that might arise from those reactions [10, 11]. The work reported in this paper concerns the relationship between cognitive states arising in a S2S-mediated map task and ASR performance, and the detection of these states by means of physiological – blood-volume pulse (BVP), and skin conductance (SC) – and speech features.

We explore the contribution of these features to the recognition of three specific user’s reactions that we considered relevant in our scenario: amusement, frustration and surprise. These states frequently occur in the setting of the Interlingual Map Task corpus used for this study and can influence the unfolding of the interaction by triggering repair strategies at different levels: at the prosodic level (variation of pitch and speaking rate), at the lexical level (production of repetition, and reformulations, change of vocabulary), at the syntactic level (use of simpler grammatical structures) [12]. The main contributions of this paper are: (1) the presentation of a new corpus of S2S interaction data with annotations for the above mentioned cognitive states which are precisely synchronised with speech signals, physiological measurements – BVP, SC, and electroencephalograms (EEG) – eye movement data, ASR and MT outputs, and video recordings, (2) a statistical analysis of the associations between cognitive states and ASR performance, and (3) an investigation on the effectiveness of cognitive state detection through machine learning in this domain by using heart rate (HR) calculated from BVP, SC, prosodic features, and combinations of these predictors.

2. Material

The Interlingual Map Task corpus consists of 15 dialogues elicited with map task technique between speakers of English and Portuguese, for a total of 9 h 40 m of speech. The data used in this paper include: high quality audio of the participants’ utterances and physiological signals (BVP, SC), which are synchronised and finely time-stamped. Orthographic transcriptions of the dialogues (using Wavesurfer [13]), and annotations using the ELAN tool [14]. In the analysed dialogues a total of
1,767 cognitive states have been annotated as shown in Table 1. Two annotators carried out the annotation of the three cognitive states: amusement, surprise and frustration. Since the annotation of cognitive states is subjective, the inter-coder agreement was calculated on one of the dialogues and the results are well above 60%.

Table 1: Summary of cog. states annotated in the ILMT corpus.

<table>
<thead>
<tr>
<th>Cognitive states</th>
<th>Count</th>
<th>Total duration (approx.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amused</td>
<td>772</td>
<td>1 h 30 m</td>
</tr>
<tr>
<td>Frustrated</td>
<td>646</td>
<td>48 m</td>
</tr>
<tr>
<td>Surprised</td>
<td>339</td>
<td>17 m</td>
</tr>
<tr>
<td>Total</td>
<td>1,767</td>
<td>2 h 35 m</td>
</tr>
</tbody>
</table>

2.1. Data collection

A prototype speech-to-speech translation system (ILMT-s2s system) shown in Figure 1, was implemented using off-the-shelf technology to enable two speakers of to communicate with each other (remotely, over the network) in their native languages [15]. The prototype enables recording of Hi-Resolution Audio (96 kHz, 24bit) of the utterances spoken to the system and stores logs and time-stamps of all relevant events performed by the system.

Figure 1: Image of the ILMT-s2s system used to collect the data

To record the biosignals, a Mind Media B.V., NeXus-4 was used to collect the BVP, SC and electrical brain activity. Additionally, a Sony HDR-XR500 camcorder was placed in front of the participants to record their movement, facial expressions and self utterances, and an SMI Eye Tracking Glasses 1.1 tracked the eye movement of the participant recording the biosignals while the other participant used a Panasonic HX-A100 to record the field of view during the whole session. An audio signal was also recorded from the biosignal device to facilitate the synchronisation of the recorded data.

2.2. Participants

Participants of the Interlingual Map Task corpus were recruited from Trinity College Dublin and also via personal network connections, with the recording taking place in a normal office environment within Trinity College Dublin. The data set consists of a total of 15 dialogues between 15 native English speakers and 15 native Portuguese speakers of both genders, aged between 18 and 45. The sessions were divided into two conditions: with video on and without video on (where the participants could or could not see each other).

2.3. Synchronisation

Synchronisation of video and audio files were performed automatically using Final Cut Pro X from Apple Inc. Videos from both subjects were synchronised and cut accordingly into one 1080p video project and output as H.264 video and audio. With all audio files aligned to the main video recording, the time difference for the ILMT-s2s system log, biosensor data time log can be determined and calculated to obtain the offset. By aligning all of the files related to the session of one subject, it becomes possible for the annotation data to be used with various data sources. For instance, annotation from the audio file can be aligned with the biosignal.

3. Experimentation

The dataset for the following experiments were created from the annotated ELAN files, which were exported as tab-delimited text files, with start time, end time, duration, annotation, transcription, ASR results and audio recording file names indicated.

3.1. Cognitive states and ASR performance

We investigate whether the three cognitive states - amusement, frustration and surprise - may arise as a consequence of errors made by the ASR system on the participant’s utterance, and whether they may in turn cause further ASR errors. Therefore we tested for associations between the word error rate (WER) and the presence of a labelled cognitive state (amusement, frustration, surprise or neutral). Our underlying hypothesis is that cognitive states that occur (immediately) before an utterance is sent for ASR, will affect ASR performance, and that ASR performance on an utterance will affect the user’s cognitive state after the user sees the ASR result. This translates into two separate null hypotheses which we seek to reject by comparing WER (our response variable) and cognitive states (the independent variable).

H1: The mean WER of utterances made after different cognitive states occur is the same for all these states.

H2: The mean WER of utterances made before different cognitive states occur is the same for all these states.

3.2. Cognitive states biosignal and prosodic features

In order to detect the cognitive states from the biosignal and the speech audio files, we compute a joint feature set for the single modalities (speech vs physiological) for which a discriminative analysis pattern classifier is tested, and then compare the results of the recognition rates for separate and integrated modalities. The idea is to verify whether the information from the biosignal combined with the prosodic analysis can improve the results of the detection.

3.2.1. Physiological features extraction

Two physiological signals are used in cognitive state recognition: The heart rate (HR) from the BVP sensor and skin conductance measured from the SG/GSR sensor. The feature set contains the median, mean, standard deviation, minimum, maximum, minimum ratio, and maximum ratio of the data values. These features are also calculated for the 1st and 2nd order derivatives of physiological signals. This results in 21 features for SC and 21 for HR. However the HR signal, maximum ratio and median of 1st and 2nd derivative have a zero value for most of the observations. These zero values are removed from the physiological feature set resulting in a total of 18 features for HR. The minimum ratio of an observation is measured by counting the number of instances which have a lower value compared to their preceding and following instance and then dividing it by the total number of instances in that observation. Similarly, the maximum ratio of an observation is measured by counting the number of instances which have a higher value compared to their preceding and following instance and then dividing it by the total number of instances in that observation.
3.2.2. Prosodic feature extraction

OpenSMILE [16] is used to extract a prosodic feature-set from the clean high-quality audio files that have been down sampled to 48 kHz, 16 bit. The feature set employed in the 2013 COMPARE challenge [17] was used. It comprises 6,373 features, including energy, spectral, cepstral, and voice-related lower-level descriptors, as well as other descriptors such as logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness.

3.2.3. Classification method

Using the MATLAB statistical tool box, a 10-fold cross-validation was performed with the discriminant analysis (DA) method for the classification of this balanced dataset. The method assumes that the feature-set of different classes follows different Gaussian distributions and follows the pseudolinear discriminative type.

3.2.4. Preprocessing and classification of dataset

The following 6 experiments were conducted to evaluate the performance of physiological and/or prosodic features for cognitive states recognition on a balanced dataset using different analysis windows as shown in figure 2. The method mentioned in 3.2.3 is used to classify the states in the below experiments. However, the prosody analysis is performed on an utterance-by-utterance basis.

**Exp. 1:** We calculate the physiological feature vectors over the 372 annotated labels of cognitive states.

**Exp. 2:** We introduce 124 observations of neutral state to the data of Exp. 1 in order to compare the physiological characteristics of neutral states with the other three states.

**Exp. 3:** Our hypothesis is that the cognitive state starts developing after the participant has read the displayed ASR result of the utterance spoken to the ILMT-s2s system. We therefore extracted 249 observations of the biosignal starting from when the ASR result is displayed on the screen and ending at the following cognitive state label end-time. We calculated the physiological feature vectors over this duration.

**Exp. 4:** Our hypothesis is that the cognitive state affect the participant’s speech even after the labelled cognitive state has ended (that is when the cognitive state is no longer observable). We selected 390 utterances spoken to the ILMT-s2s system that occurred after the labelled cognitive state and calculated the prosodic feature vectors for these utterances.

**Exp. 5:** We investigate whether the cognitive state will affect the participant’s physiological characteristics even after the cognitive state is no longer observable (after the labelled section has ended). We calculated the physiological feature vectors for the 222 observations over an extended window starting when the labelled section starts, but ending when the following speech utterance ends.

**Exp. 6:** We combine the physiological features from the fifth experiment with the prosodic feature of the utterances that followed the cognitive state (on the time-line) of the ILMT-s2s system.

4. Results and discussions

4.1. Cognitive states and ASR performance

For the hypotheses in 3.1 the following results were obtained. See Figure 3 for details.

**H1:** No statistically significant difference (Analysis of variance, \( F_{3,28} = 1.309; \ p = 0.251 \)) was observed between the cognitive state before the utterance that was sent to the ASR system and the resulting ASR WER.

**H2:** A statistically significant difference \( F_{3,28} = 9.671; \ p < 0.001 \) was observed between the cognitive state after utterances sent to the ASR system and the resulting ASR WER of that utterance.

For H2, Post-hoc comparisons (Tukey HSD test) revealed significant differences between surprise and neutral \( (p < 0.01) \), and frustration and neutral \( (p < 0.01) \). Differences between frustration and amusement and surprise and amusement were also found \( (p < 0.05) \). The strongest reaction to ASR errors was surprise (mean WER 46.4%), followed by frustration (41.9%).

**Figure 2:** Analysis window explanation.

**Figure 3:** Graphic representation of the ASR WER scores.

4.2. Cognitive state recognition

Table 2 and Figure 4 show the recognition rates of the classification. The average rates of classification for physiological data (combination of SC and HR) in the 1st experiment is 61.29% and improves in the 2nd experiment to 70.77%. However, in the 3rd experiment, the average percentages of classification accu-
Figure 4: The figure shows the classifier average accuracies.

The results here reported validate the hypothesis that the performance of bimodal cognitive state recognition gives better results compared to unimodal recognition: the bimodal approach (combination of prosodic and physiological characteristics provides 68.02% accuracy) gives an improvement of almost 7% compared to the performance of the physiological signals (61.29%). The results of physiological signals using neutral labels show that the characteristics of neutral labels are clearly separable from the three cognitive states, which resulted in good classification of these labels. Moreover, these results also show that the HR of neutral state is completely different from the three cognitive states. In addition it is observed that HR has more correlation with amusement than SC, while, SC is more correlated with frustration and surprise than HR. Using the “extended window” results in an increase of accuracy for amusement but a decrease in frustration and surprise accuracies.

### 4.3. Future work

A statistically significant difference is observed in the ASR WER of utterances that are followed by a cognitive state (H2), whereas no differences is found with respect to WER of utterances produced after the onset of a cognitive state. However, the participants’ voice still contained prosodic features of the following cognitive states that were automatically identifiable approximately 50% of the time. It will be interesting to see if it is possible to obtain a statistically significant difference for instances that are correctly automatically identifiable. In addition to utterances sent to ASR, our ILMT-s2s corpus also contains a spontaneous speech track (where the user “speaks aloud” to themselves). In future we plan to assess the effectiveness of speech features extracted from this spontaneous speech track in detecting cognitive states.

### 5. Conclusions

An investigation of the possible associations between speech recognition performance and three cognitive states (amusement, frustration and surprise) that arise in dialogues mediated by a speech-to-speech machine translation system is here presented. The results indicate that (a) these cognitive states often arise as a consequence of what happens in the Speech-to-Speech mediated interaction with a statistically significant difference obtained between the WER of an utterance and the different cognitive state after the utterance, (b) the association between the cognitive state and the biosignals does not seem to persist until the next sentence is uttered, as suggested by the poor state detection performance in time windows that include following utterance, and by the statistical testing, and (c) that features of the speech signal can be used to complement biosignals in detecting cognitive states in time windows that include the following utterance. Extending the window to the end of the utterance following the cognitive state yields poor detection on biosignals alone, but improves considerably if features of the speech signal are added, thus showing the potential usefulness of speech features as a biosignal.

### 6. Acknowledgements

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/1267) in the ADAPT Centre (www.adapcentre.ie) at Trinity College Dublin, and FP7-METALOGUE project under Grant No. 611073.
7. References


