Facial expression as an input annotation modality for affective speech-to-speech translation

Éva Székely, Zeeshan Ahmed, Ingmar Steiner, and Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics, University College Dublin

{eva.szekely,zeeshan.ahmed}@ucdconnect.ie
{ingmar.steiner,julie.berndsen}@ucd.ie

Abstract One of the challenges of speech-to-speech translation is to accurately preserve the paralinguistic information in the speaker's message. In this work we explore the use of automatic facial expression analysis as an input annotation modality to transfer paralinguistic information at a symbolic level from input to output in speech-to-speech translation. To evaluate the feasibility of this approach, a prototype system, FEAST (Facial Expression-based Affective Speech Translation) has been developed. FEAST classifies the emotional state of the user and uses it to render the translated output in an appropriate voice style, using expressive speech synthesis.

1 Introduction

There has been a considerable recent interest in expressive speech synthesis systems. The ability of a speech synthesiser to express emotion and affect undoubtedly improves its ability to facilitate spoken communication. Speech-to-speech translation is an application where speech synthesis is used as a communication tool between humans, making the appropriate expression of affect in the synthetic speech all the more important [2]. Unlike other applications such as text-to-speech (TTS) systems, where affect and emotion would need to be predicted from the textual input of the synthesiser, speech-to-speech translation systems can apply processing strategies to multimodal input, to classify and output the paralinguistic information in a speaker's intended message.

Agüero et al. [1] proposed a method to preserve the paralinguistic information of the input speech in the translated synthetic output speech by transmitting f0 contours across Spanish and Catalan speech. While "transplanting" prosody based on acoustic information may produce satisfactory results for closely related language pairs, this is unlikely to be the case when translating across languages that are very different.

In this work, we take a different approach by transporting paralinguistic information at a symbolic level from visual input to acoustic output. Essentially, the idea is to automatically analyse the facial expression of the speaker, and process this interpretation as paralinguistic information alongside the speech translation, by mapping the underlying emotion of the speaker's facial expression to the voice style of a speech synthesiser. The speaker's emotional state interpreted from his facial expression is transferred as an abstract concept in a paralinguistic analogy of "interlingual transposition" [7], i.e., the translation from the source to the target by means of an intermediate high-level representation.

Previous studies which processed multimodal input (face and voice) for emotion recognition have reported promising results [5, 13]. For the purposes of affective speech

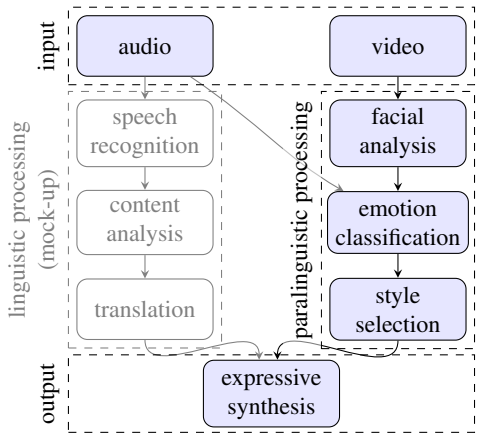

Figure 1. System architecture of FEAST

translation, it is desirable to apply a method to recognise the emotional state of the speaker which is as language-independent as possible. While the expression of emotion through facial features may show some differences across cultures, visual expressions of emotion are likely to be less language-dependent than vocally expressed emotional features [6].

The goal of this study is to assess the extent to which this preservation of the speaker's paralinguistic (implicit) message is possible based on analysing visual input alone. In order to test this, the FEAST prototype system has been developed, focusing on the task of recognising and preserving "stereotypical" representations of three basic emotions, *happy*, *sad*, and *angry* (or emotionally neutral input), at the utterance level. The output of the system is generated by an expressive speech synthesiser that includes voice styles reflecting each of these emotional states. The extension of the system to process more nuanced expressions of affect through dimensional approaches, as well as the integration of acoustic features of emotion to improve classification accuracy, is a subject of future work.

2 System architecture and processing workflow

The FEAST prototype system takes multimodal input in the form of video and audio, processes the linguistic and paralinguistic aspects in tandem, and generates spoken output by means of a speech synthesiser. A diagram of the system architecture is shown in Figure 1.

The linguistic content is extracted from the input audio using automatic speech recognition (ASR) and automatically translated into the target language. For the time being, these components are implemented only as mock-ups to simulate the functionality of the final system.

On the paralinguistic side, the video input is processed by a face detection and analysis component, which extracts the facial expression of the speaker from the video frames. The resulting features are subsequently classified into emotion categories, which are then used to select an appropriate synthesis style.

The speech synthesiser as the final component takes as input the textual representation of the linguistic content, as well as the voice style selected by the paralinguistic processing, and generates the spoken translation rendered in the appropriate style.

3 Linguistic processing

The current version of the system focuses on identifying and preserving the paralinguistic information of the audiovisual input in the translated synthetic speech. In order to reliably evaluate this function of the system without having to deal with noise or errors introduced by automatic speech recognition and machine translation systems, these two components of the FEAST system are assumed to be working perfectly, and are simulated by mock components.

For the purposes of demonstration and evaluation, a set of 24 utterances was prepared containing both the transcription of English input utterances as well as their German translation as produced by a human translator.

4 Paralinguistic processing

The system components which process paralinguistic features comprise individual components for face detection and analysis (Section 4.1), emotion classification (Section 4.2), and style selection (Section 4.3).

4.1 Face detection and analysis

The face detection and expression analysis used in this study is performed by the SHORE library for real-time face detection and fine analysis.[1] An application programming interface (API) for the system has been made available by Fraunhofer Institute for Integrated Circuits IIS for academic demonstration and evaluation purposes. When detecting faces and facial expressions, SHORE analyses local structure features in an image (or series of images) that are computed with a modified census transform [8]. This face detection system outputs scores for four distinct facial expressions, *angry*, *happy*, *sad*, and *surprised*, with a value for the intensity of the expression, as well as a confidence measure. If a face is detected in an image with no facial expression values, it can be interpreted as a *neutral* face.

The SHORE library has previously been integrated with an English language expressive speech synthesiser for an application developed for use in speech generating devices of non-speaking individuals [12], where static images were processed for utterance production. Because the free version of the SHORE API can analyse still images in real-time, for the purposes of this study, the API was adapted for frame-by-frame video analysis, using the OpenCV platform.[2]

4.2 Emotion classification

The aim of the facial expression analysis in FEAST is to output a *single* decision regarding the emotional state of the user over each utterance. To optimise the performance for utterance-level analysis, and in particular to deal with the fact that the user is speaking (which changes the facial expression from frame to frame, especially wrt. the shape of

[1] http://www.iis.fraunhofer.de/en/bf/bsy/fue/isyst/
[2] http://opencv.org/

the lips), the training of a visual emotion classifier was deemed necessary. This classifier was trained on selected segments of the SEMAINE database [9]. Details of the classifier training and evaluation are given in Section 5.1.

4.3 Style selection

After the emotional state of the speaker has been classified, the style for the expressive speech synthesiser is determined or selected from a list of available styles. In the current prototype, this amounts to a straightforward mapping from emotion to voice style. Videos classified as *happy* are synthesised with *cheerful* style, *sad* with *depressed*, and *angry* with *aggressive*. If the speaker's affective state is classified as *neutral*, the speech translation results in a *neutral* voice style.

For future extensions of FEAST involving dimensional representations of emotion, this component could be responsible for more sophisticated voice style control.

4.4 Expressive speech synthesis

The TTS component uses the open-source synthesis platform MARY [10].[3] MARY provides language resources and voices for a number of languages, including German, as well as engines for diphone, unit-selection, and hidden Markov model (HMM)-based synthesis.

For expressive unit-selection synthesis, MARY includes facilities to select units based on appropriate symbolic or acoustic features [11]. A male German unit-selection voice named `dfki-pavoque-styles` which incorporates this feature is available;[4] it contains data from a single-speaker, multi-style speech corpus, and allows TTS requests to specify either *cheerful*, *depressed*, or *aggressive* speaking style, in addition to the default *neutral* style.

In this component of FEAST, the textual representation of the translated content is wrapped into an HTTP request for processing by the MARY TTS server. The classification result of the emotion analysis component is mapped onto one of the expressive styles available in the `dfki-pavoque-styles` voice, which is added to the HTTP request as a `STYLE` parameter.

The resulting synthesis request is then processed by the TTS server, producing a WAV file which is then played back locally to the user.

5 Evaluation

If we hypothesise that the preservation of the emotion through expressive synthetic speech improves listeners' experience of speech-to-speech translation, several questions need to be answered to evaluate the performance of the system and its individual components:

(1) Does the system accurately classify emotion on the utterance level, based on the facial expression in the video input?
(2) Do the synthetic voice styles succeed in conveying the target emotion category?
(3) Do listeners agree with the cross-lingual transfer of paralinguistic information from the multimodal stimuli to the expressive synthetic output?

Three evaluation experiments were conducted to address these questions.

[3] `http://mary.dfki.de/`
[4] Released under the Creative Commons Attribution-NoDerivatives license.

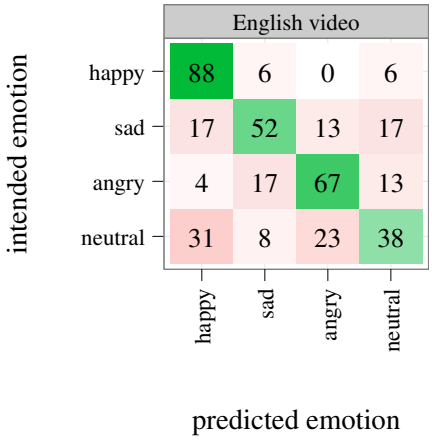

Figure 2. Results of the emotional state classification for video. Cell shading indicates correct (green) vs. incorrect (red) classification.

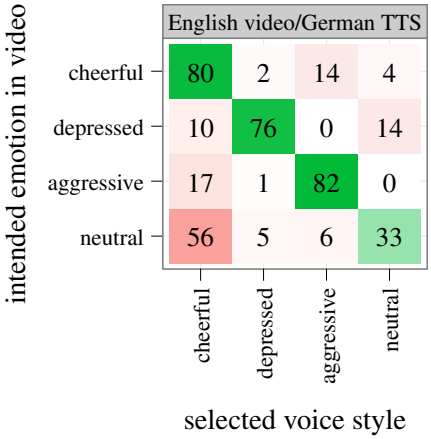

Figure 3. Results of the perceptual test comparing audiovisual input with translated audio output.

5.1 Classification of emotion from facial expression

As mentioned earlier, the classifier used in the emotion classification component was trained on the SEMAINE database [9]. This database was recorded to study natural social signals that occur in (English) conversations between humans and artificially intelligent agents, and to collect video data that could be used for the training of such agents.[5] For the recordings, the participants were asked to interact with four emotionally stereotyped characters portrayed by an actor. These characters are Poppy, who is happy and outgoing; Obadiah, who is sad and depressive; Spike, who is angry and confrontational; and Prudence, who is even tempered and sensible.

For the training of the classifier, we selected the video recordings of the male operators in the SEMAINE database: a set of 642 utterances was extracted from the video database and each video frame was analysed using SHORE. The character played by the actors in these video sequences can be used as a positive classifier for the example data. Ideally, the utterances for Poppy should be classified as *happy*, Obadiah as *sad*, Spike as *angry*, and Prudence as *neutral*, based on the facial expression analysis.

From the SHORE analysis on each frame, the following features were extracted to build a support vector machine (SVM) classifier: average feature value for each facial expression, the 20th, 50th and 90th percentile of these values and the percentage of frames capturing each expression or a neutral expression. We trained a SVM with a Radial Basis Function (RBF) kernel on 5/6 of the sentences extracted from the videos (535 utterances). The classifier was implemented using the LIBSVM software system [4].[6] Optimal parameters for the RBF kernel and the relevant features were selected using a grid search and 5-fold cross validation on the training data.

[5] The database is freely available for scientific research purposes at http://semaine-db.eu/.
[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

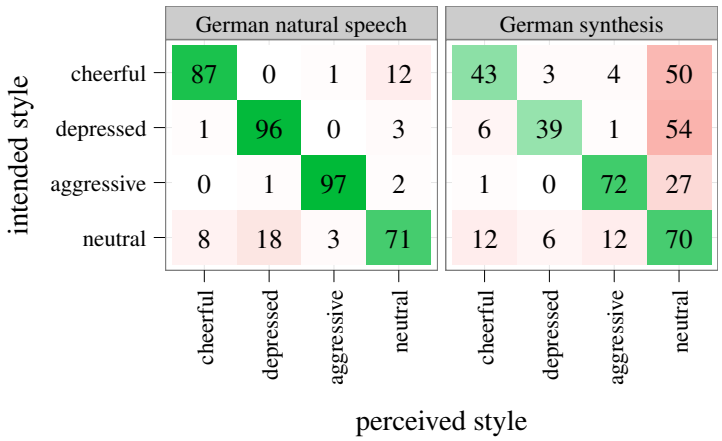

Figure 4. Contingency table of identification task results for intended vs. perceived voice style, for original recordings (left) and expressive unit-selection synthesis with a mixed-style voice (right).

Using this model on the test data (107 utterances) an accuracy of 63.5 % was achieved ($F1 = 65.1$). Figure 2 presents the results of the classification for each emotional state.

5.2 Perception of style in expressive synthesis

To assess whether the expressive voice styles in the `dfki-pavoque-styles` voice data are perceived as intended, and how this perception is affected by mixed-style unit-selection synthesis, a perception experiment was conducted. Five sentences of neutral content were selected from the corpus, each spoken in a *cheerful*, *depressed*, *aggressive*, and *neutral* style. In addition, the sentences were synthesised in each of these voice styles, using MARY with a mixed-style voice containing both neutral and expressive units; prosody was predicted by classification and regression trees (CARTs) [3] trained only on the corresponding subset of the corpus.

A group of 20 native speakers of German (undergraduate university students, 11 f/9 m) was recruited as a pool of paid subjects for the experiment. Each subject was asked to listen to the original and synthesised stimuli and identify which of the four voice styles best described each one; the response categories were *cheerful*, *depressed*, *aggressive*, and "none of these". Using Praat and its "ExperimentMFC" facility,[7] the stimuli were presented in randomised order over headphones in a quiet environment. The results of the style identification task are given in Figure 4.

5.3 Perception of paralinguistic adequacy for speech-to-speech translation

To evaluate the adequacy of the symbolic, cross-lingual transfer of paralinguistic information from the multimodal stimuli to the expressive synthetic output, a third experiment was conducted. The evaluation was implemented using a password-protected webpage.

[7] `http://praat.org/`; Multiple forced choice listening experiment described at `http://www.fon.hum.uva.nl/praat/manual/ExperimentMFC.html`.

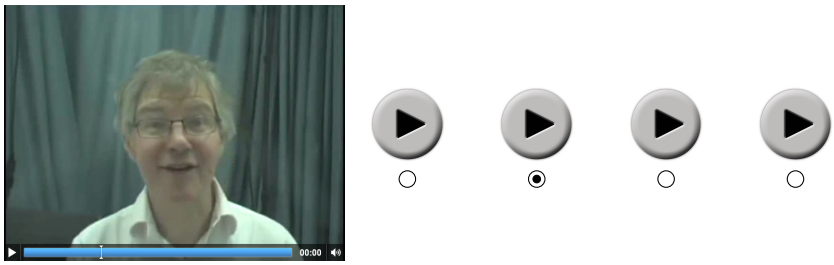

Figure 5. Example of one stimulus from the perception experiment. At left, the video of the English utterance, followed by buttons to play audio samples of the German translation, synthesised in each of four different voice styles. The subject's task is to select the radio button below the audio sample which best conveys the emotion portrayed by the speaker in the video.

For this evaluation, 24 utterances were selected from the recordings of one male operator from the SEMAINE database, 6 for each character type (Poppy, Obadiah, Spike, and Prudence). After reading a short introduction, the participants were asked to play the English videos and select from the four available expressively synthesised renditions of the German translation the one which they felt was the best match for the original emotion portrayed in the video (cf. Figure 5). The order of the videos as well as the order of the corresponding synthetic samples were randomised for each trial.

The subjective listening test was carried out by 14 bilingual participants, 5 of them native speakers of German. All participants had a good comprehension of English. The results are summarised in Figure 3.

6 Discussion

Because of the small sample size possible to evaluate with a perceptual test, it is difficult to tell exactly what percentage of the classified output and matched voice style listeners would agree with. The reason for this is that the error potential of the system is two-stage: a video may be classified incorrectly, or a particular correctly classified video may not match the mapped voice style according to a listener. If FEAST is being used in a real-life situation, it is necessary to weigh the type of classification errors. Hereby, classification errors across emotions should be avoided at the cost of classification of an emotional state as *neutral*. This can be done through only processing the classification outputs where the classifier's confidence is high, for the rest of the utterances, the system would stay on the "safe side", and synthesise the output with a neutral voice style. That said, it is reasonable to think that even a small percentage of correctly identified and transferred emotional state could result in significant improvement of user's experience with a speech-to-speech translation system.

7 Conclusion and future work

The evaluation has demonstrated on examples of speech translation from English speaking videos to German synthetic speech output that preserving the intended paralinguistic content of a message is possible with significantly greater than chance accuracy, when considering distinct categories of three basic emotions, and the neutral emotional state. Our language-independent classifier based on facial expressions identified emotional

state with an overall 63.5 % accuracy, with the emotions *happy* and *angry* being more easily classifiable than *sad* and *neutral*. It becomes apparent in all three evaluations that *cheerful/happy* is often mistaken for *neutral*. However, from a usability perspective this is much more acceptable than systematic confusion of either with negative affect.

The classifier will be extended with the capability to take multimodal input to compute the prediction of the affective state of the user based on acoustic and prosodic analysis as well as facial expressions. Future work further involves developing a demonstration of the prototype system that takes live input through a webcamera and microphone. The integration of a speech recogniser and a machine translation component for a fully functional affective speech translation is planned.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (http://cngl.ie/) at University College Dublin (UCD) and Trinity College Dublin (TCD). The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland. Portions of the research in this paper use the Semaine Database collected for the Semaine project (www.semaine-db.eu) [9].

Bibliography

[1] Agüero, P.D., Adell, J., Bonafonte, A.: Prosody generation for speech-to-speech translation. In: Int. Conf. Acoust. Speech Signal Process. pp. I–557–560 (2006)

[2] Batliner, A., *et al.*.: The recognition of emotion. In: Wahlster, W. (ed.) Verbmobil: Foundations of Speech-to-Speech Translations, pp. 122–130. Springer (2000)

[3] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth (1984)

[4] Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Tech. 2(3), 27:1–27:27 (2011)

[5] Cowie, R., *et al.*.: Emotion recognition in human-computer interaction. Signal Process. Mag. 18(1), 32–80 (2001)

[6] Ekman, P., Keltner, D.: Universal facial expressions of emotion: an old controversy and new findings. In: Segerstråle, U., Molnár, P. (eds.) Nonverbal Communication: Where Nature Meets Culture, pp. 27–46. Lawrence Erlbaum (1997)

[7] Jakobson, R.: On linguistic aspects of translation. In: Brower, R.A. (ed.) On Translation, pp. 232–239. Harvard University Press (1959)

[8] Küblbeck, C., Ernst, A.: Face detection and tracking in video sequences using the modified census transformation. Image Vision Comput. 24(6), 564–572 (2006)

[9] McKeown, G., Valstar, M.F., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: Int. Conf. Multimedia Expo. pp. 1079–1084 (2010)

[10] Schröder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. Int. J. Speech Tech. 6(4), 365–377 (2003)

[11] Steiner, I., Schröder, M., Charfuelan, M., Klepp, A.: Symbolic vs. acoustics-based style control for expressive unit selection. In: ISCA Wkshp. Speech Synth. pp. 114–119 (2010)

[12] Székely, É., Ahmed, Z., Cabral, J.P., Carson-Berndsen, J.: WinkTalk: a demonstration of a multimodal speech synthesis platform linking facial expressions to expressive synthetic voices. In: Wkshp. Speech Lang. Process. Assist. Tech. pp. 5–8 (2012)

[13] Wöllmer, M., *et al.*.: Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In: Interspeech. pp. 2362–2365 (2010)