J Multimodal User Interfaces (2014) 8:87–96
doi:10.1007/s12193-013-0128-x

Facial Expression-based Affective Speech Translation

Éva Székely · Ingmar Steiner · Zeeshan Ahmed · Julie Carson-Berndsen

Received: 7 April 2013 / Accepted: 8 July 2013 / Published online: 30 July 2013

Abstract One of the challenges of speech-to-speech translation is to accurately preserve the paralinguistic information in the speaker's message. Information about affect and emotional intent of a speaker are often carried in more than one modality. For this reason, the possibility of multimodal interaction with the system and the conversation partner may greatly increase the likelihood of a successful and gratifying communication process. 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.

Electronic supplementary material The online version of this article (doi:10.1007/s12193-013-0128-x) contains supplementary material, which is available to authorized users.

É. Székely (✉) · Z. Ahmed · J. Carson-Berndsen
Centre for Next Generation Localisation, School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland
E-mail: eva.szekely@ucdconnect.ie

Z. Ahmed
E-mail: zeeshan.ahmed@ucdconnect.ie

J. Carson-Berndsen
E-mail: julie.berndsen@ucd.ie

I. Steiner
Multimodal Computing and Interaction, Saarland University and Language Technology Lab, DFKI GmbH, Campus C7.4, 66123 Saarbrücken, Germany
E-mail: steiner@coli.uni-saarland.de

Keywords Expressive speech synthesis · Speech-to-speech translation · Gesture-driven multimodal interface · Affective computing

1 Introduction

Speech-to-speech translation is an application where speech recognition, machine translation, and speech synthesis are used together as a communication tool between humans speaking different languages. Where human-to-human communication is mediated by a machine, the mere processing of linguistic content is insufficient to guarantee communication success. To ensure seamless understanding between conversation partners, the mediator system needs to be able to detect and transmit paralinguistic, affective information in real time. Information about the emotional state of the speaker, as well as affective nuances in the intent of a message are often contained in more than one modality: voice, facial expression, hand gestures, posture, etc. The real time processing of such multimodal information not only makes the process of transmission more robust, but also opens the door to exploiting communication elements that are less language dependent, simultaneously with the translation of linguistic content between two languages. The role of the speech synthesiser in expressing emotion and affect is crucially important in this process [3]. 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 reflect the paralinguistic information from a speaker's intended message.

The primary input required for a successful speech-to-speech translation process, is the linguistic content of the user's speech captured by a speech recogniser. However, many speech-to-speech translation systems have benefited

from integrating additional sources of information, with the aim of improving the accuracy of the recognition task [19], enhancing the user's experience with the system through a multimodal interface [16], or influencing the synthetic speech output to be – in some aspect – more similar to the source speech signal. Within the latter task, the main target of research effort has been to preserve the identity of the speaker in the target language. The main approaches to this include cross-lingual speech synthesis and voice conversion techniques [11].

A less prevalent, yet emerging focus of interest is to transmit paralinguistic information from the source to the target speech, in order to better capture the nuances of the input message and minimise the chance of misunderstandings due to incorrect representation of prosodic and emotional features.

Agüero et al. [1] aim to preserve the prosody of the input speech in the translated synthetic output speech by transmitting F_0 contours between Spanish and Catalan speech. While this method may produce good results for closely related language pairs, when translating across languages that are very different, a less language dependent approach might be desirable.

Kano et al. [9] propose a language independent method to translate paralinguistic information from source to target speech by transferring acoustic features such as duration and power to the output speech. This method is able to transmit information about emphasised words in sentences, which is useful in situations where a message needs to be repeated because of a previous mistake, so that the word where the mistake was made can be emphasised in the target language as well.

In the present paper, another type of paralinguistic information is targeted, namely information carried about the emotional state of the user. As a first step towards a method that translates affective states independently of the source language, we aim to integrate visual sources of information into the speech-to-speech translation process, by using facial expression as an input annotation modality. 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 or her facial expression is transferred as an abstract concept in a paralinguistic analogy of interlingual transfer [20], 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 [7, 21]. For the purposes of affective speech 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 [8].

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. The speech-to-speech translation component is implemented using the

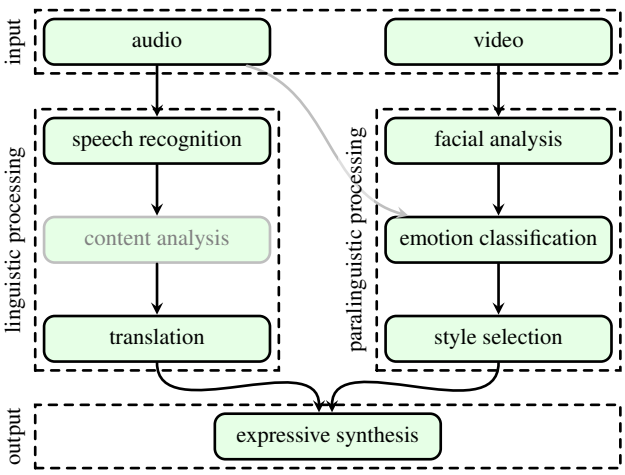

Fig. 1 System architecture of FEAST. The content and audio are not yet used for emotion classification.

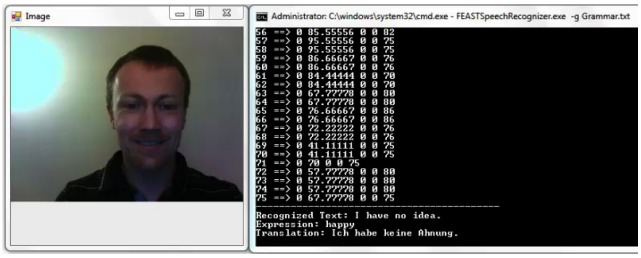

Fig. 2 Screenshot of the FEAST system in action [from 2]). The camera captures the user's face and displays it in the left window, while the console window to the right logs the recognised utterance and its translation, along with the facial expression classification results. The translation is synthesised using the appropriate expressive speaking style and played back (not shown).

Microsoft Speech software development kit (SDK)[1] and the Bing Translation application programming interface (API).[2]

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 MARY TTS synthesiser [15] 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.

A short video presenting the functionalities of the FEAST demo system can be viewed in Online Resource 1; a screenshot is shown in Figure 2.

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. The linguistic processing can comprise speech recognition, content analysis, and machine translation components, as shown on the left side of Figure 1. Currently, the FEAST system predicts the emotion for the target speech using only the facial expression from the video input. However, the system also provides the scope for integrating additional components, e.g., audio and text content analysis, which could contribute to predicting emotion for the speech output.

The speech recognition component for FEAST is implemented using the Microsoft Speech SDK that provides general-purpose acoustic models for English ASR. For better accuracy, we restrict the recognition to the application domain. The translation component is based on the Bing Translation API that provides the German translation of the given English input.

The translated text is then combined with the affective state determined by the facial expression analysis components to form an Emotion Markup Language (EmotionML) document.[3] Finally, this is sent to the synthesis server, yielding the translation output, spoken in the target style.

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 4.4).

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.[4] An API for the system has been made available by Fraunhofer Institute for Integrated Circuits 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 [10]. Hereby the images are summarised by local feature kernels, 3×3 images where the image local structure is summarised in binary notation. These feature kernels allow for fast identification of facial features such as the eye and mouth distance, as well as the curve of the mouth and whether it or the eyes are open.

Using the ImageMarker application of the SHORE toolbox, analysed and identified images are annotated with additional information such as gender, facial expression and age. Various classifiers are trained on this data, and combined into one strong classifier using the AdaBoost method [10]. The fine analysis 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. The latter two range from 0 to 100, with a higher value indicating higher intensity and likelihood, respectively. 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 [18], where static images were processed for utterance production. Because the SHORE API can analyse still images in real-time, for the purposes

[1] http://www.microsoft.com/en-us/download/details.aspx?id=10121

[2] http://www.microsofttranslator.com/dev/

[3] http://www.w3.org/TR/emotionml/

[4] http://www.iis.fraunhofer.de/shore

of this study, the API was adapted for frame-by-frame video analysis, using the OpenCV platform.[5]

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 the lips), the training of a visual emotion classifier was deemed necessary. This classifier was trained on selected segments of the SEMAINE database [12]. Details of the classifier training and evaluation are given in Section 5.1.

4.3 Real-time emotion classification on video

The SHORE library provides facilities to analyse facial expressions in static images. To identify the affective state of a speaker in a video, the video frames are first analysed individually, and all possible emotion categories are generated with their confidence scores within each frame. The system then takes the average of each score over all frames and classifies the video with the emotion category that receives the highest score. A snapshot of the running system is shown in Figure 2, with scores for each emotion category, as well as recognised and translated text, displayed in the console window.

4.4 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. Utterances 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.

4.5 Expressive speech synthesis

The TTS component uses the open-source synthesis platform MARY [15].[6] 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. MARY also supports the input to be specified using EmotionML for expressive speech synthesis [14].

```
1  <emotionml version="1.0" xmlns="http://www.w3.org...
       /2009/10/emotionml" category-set="http://www.w3...
       .org/TR/emotion-voc/xml#everyday-categories">
2    <emotion>
3      <category name="happy"/>
4      Haben Sie schon einen Termin?
5    </emotion>
6  </emotionml>
```

Listing 1 Example EmotionML document used as input for the expressive TTS; the speaking style is controlled using the category name attribute (line 3), viz., *angry*, *happy*, or *sad*.

For expressive unit-selection synthesis, MARY includes facilities to select units based on appropriate symbolic or acoustic features [17]. A male German unit-selection voice which incorporates this feature is available;[7] 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 EmotionML document for processing by the MARY TTS server. The classification result of the affective state analysis component is mapped to one of the expressive styles available in the synthesis voice, which is added to the EmotionML document as an "everyday" category name [6].[8] An example EmotionML document is shown in Listing 1.

The resulting synthesis request is then processed by the TTS server, producing an audio file which is then played back 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?

[5] http://opencv.org/

[6] http://mary.dfki.de/

[7] `dfki-pavoque-styles`, released under the Creative Commons Attribution-NoDerivatives 3.0 license.

[8] http://www.w3.org/TR/emotion-voc/

3. Do listeners agree with the cross-lingual transfer of paralinguistic information from the multimodal stimuli to the expressive synthetic output?
4. How is the overall performance of the system, including the interaction of machine translation and voice style selection?

A number of evaluation experiments were conducted to address these questions.

5.1 Classification of emotion from facial expression

To assess the potential of utterance-level emotion classification based on facial expressions on videos of a person talking, a classifier was trained on the SEMAINE database [12]. 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.[9] 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 [5].[10] 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.

Using this model on the test data (107 utterances) an accuracy of 63.5 % was achieved ($F1 = 65.1$). Figure 3 presents the results of the classification for each emotional state.

[9] The database is freely available for scientific research purposes at http://semaine-db.eu/.

[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

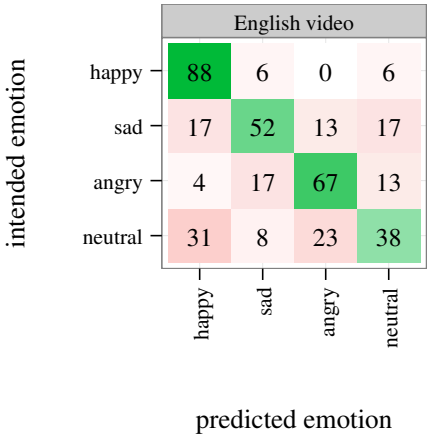

Fig. 3 Results of the emotional state classification for video. Cell shading indicates correct (green) vs. incorrect (red) classification.

emotion	number of utterances	correct	
		raw	adapted
angry	148	7	52
happy	202	195	190
neutral	139	29	23
sad	148	0	5
total	637	231	270

Table 1 Statistics of SEMAINE corpus subset selected for evaluation, and accuracy of FEAST system, before ("raw") and after adaptation.

5.2 Evaluation of real-time automatic voice selection

The FEAST system currently uses a classifier that is based on average facial expression scores per utterance, and works in real-time (see Section 4.3). An automatic evaluation of this classifier was performed on a subset of the SEMAINE corpus [12] using the annotation provided in the corpus.

The purpose of the automatic evaluation is to highlight how accurately this method classifies the human emotion portrayed in the video, and to provide a comparison to a data-driven classification method (described in Section 5.1) which, however, does not provide real-time output.

Table 1 shows the statistics for the data selected for evaluation, as well as the results of running the FEAST system on that data. We selected 637 utterances from two male operators, out of which 148 utterances were manually classified as spoken in an *angry* style, 202 as *happy*, 148 as *sad*, and 139 as *neutral*.

According to the results presented in Table 1, the overall FEAST system performance does not seem encouraging on the SEMAINE corpus; the overall accuracy is 36.26 %. Looking more closely at each emotion category, we find that the system performs very well on *happy* utterances, while *sad* and *angry* utterances are almost never correctly clas-

sified. One likely explanation for this is that for facial expression classification, we use the off-the-shelf SHORE library, which was trained on still faces rather than talking faces. Another reason for the low performance is the mismatch between the training and test data environments. Addressing the former problem would require the system to be completely retrained on talking faces. The latter problem can be reduced using the adaptation strategy discussed in the following section, which results in considerable improvement.

5.2.1 Adaptation

We try to compensate for the training and test data mismatch problem by applying a weight to the score of each emotion category classified by SHORE. Currently, the system chooses an emotion category that is scored highest by SHORE, i.e.

$$\hat{x}_f = max_f(x_1, x_2 \ldots x_n) \tag{1}$$

where x_i is a score of an emotion category i generated by SHORE for a video frame f and $\hat{x}_f$ is the emotion category that has the maximum score among all the emotion categories for the frame f. When applying weights to each emotion category, the selection of the prevalent emotion in the video frame is performed as follows.

$$\hat{x}_f = max_f(a_1 x_1, a_2 x_2 \ldots a_n x_n) \tag{2}$$

where a_i is a weight for emotion category i. The weights are learned using a standard machine learning approach, minimum error rate training (MERT) [13]. In this approach, the weights are chosen in a way that minimises the decision error rate on the development set.

It is obvious from the results reported in Table 1 that the system is strongly biased toward recognising *happy* emotion in the SEMAINE data. After adjusting weights on a separate development set, we obtained the improved accuracy, as shown in Table 1 and Figure 4.

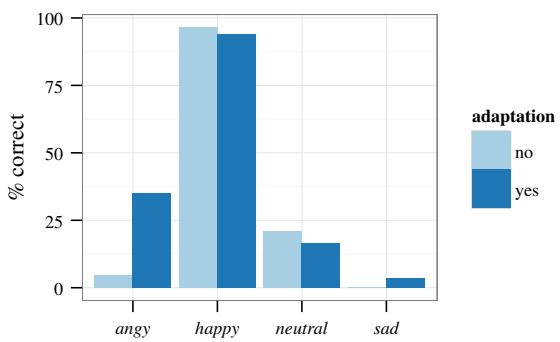

Fig. 4 Correct classification of utterances by portrayed emotion in SEMAINE corpus, before (light) and after (dark) adaptation.

The adapted FEAST system shows improved performance. The overall system accuracy is 42.38 %, which is 16.87 % better than the unadapted baseline. Facial expression classification for *angry* improves considerably, while *sad*, which was originally not recognised at all, receives some improvement.

5.3 Perception of style in expressive synthesis

To assess whether the expressive styles in the 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 SEMAINE 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) [4] 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,[11] the stimuli were presented in randomised order over headphones in a quiet environment. The results of the style identification task are given in Figure 5.

5.4 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, another experiment was conducted. The evaluation was implemented using a password-protected webpage.

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, cf. Section 5.1). For the purposes of the evaluation, the German translation of these 24 utterances was produced by a human translator. This was done in order to evaluate voice style selection alone, without the interference of possible linguistic errors in the output due to imperfect machine translation. The interference of both automatic components (voice selection and MT) is evaluated

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

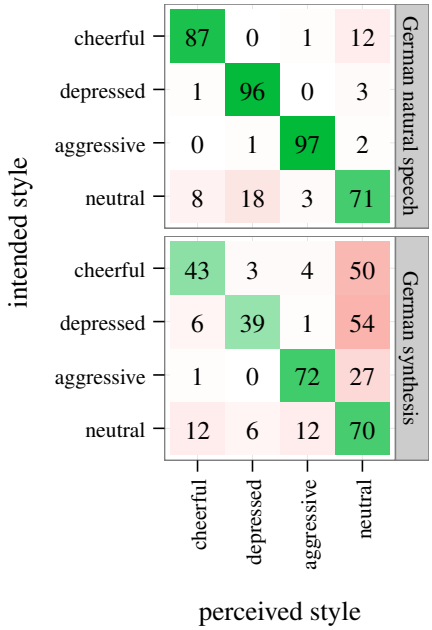

Fig. 5 Contingency table of identification task results for intended vs. perceived voice style, for original recordings (top) and expressive unit-selection synthesis with a mixed-style voice (bottom).

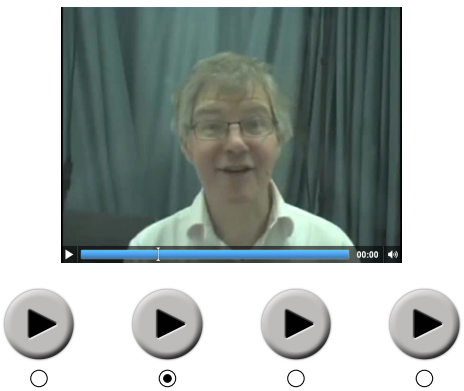

Fig. 6 Example of one stimulus from the perception experiment. Above, the video of the English utterance [from 12], followed by buttons to play audio samples of the German translation, synthesised in each of four different voice styles (in randomised order). The subject's task is to select the radio button below the audio sample which best conveys the emotion portrayed by the speaker in the video.

in Section 5.5. 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 6). The order of the videos as well as the order of the corresponding synthetic samples were randomised for each trial.

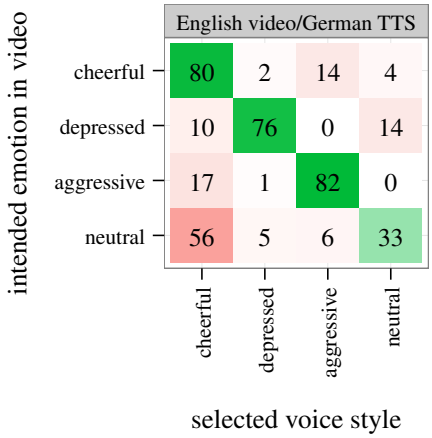

Fig. 7 Results of the perceptual test comparing audiovisual input with translated audio output.

The subjective listening test was carried out by 14 participants, 5 of them native speakers of German. All participants had a good comprehension of English. The results are summarised in Figure 7.

5.5 Evaluation of affective speech translation

Finally, an online evaluation of the system performance was conducted in order to investigate the effect of the voice style choice on the understanding of the translated message by the listener. Specifically, we were interested to see how the combination of facial gesture-directed expressive speech synthesis and machine translated text affects the comprehension of the message, and whether a suitable voice style may help avoid misunderstandings in cases where there are minor errors in the translation.

For this evaluation, we used video segments from acted conversations that contain emotional content but where – unlike in the SEMAINE database – the subject was not instructed to perform a particular stereotype of a basic emotion, merely to act out the prompts as he saw fit. This was to create a more realistic scenario to a real life speech translation task, and assess the ability of the four voice styles to cover the varied emotional intent represented in dialogues. Each of the 20 video segments used in this online evaluation contain one utterance and feature a 28 year-old male native speaker of (Irish) English.

For each trial, evaluation participants were asked to watch the video segment, and then listen to the translated German synthetic speech in the voice style selected by the FEAST system, before answering four questions:

I. whether the speaker's intention would have been misunderstood in the absence of visual input;

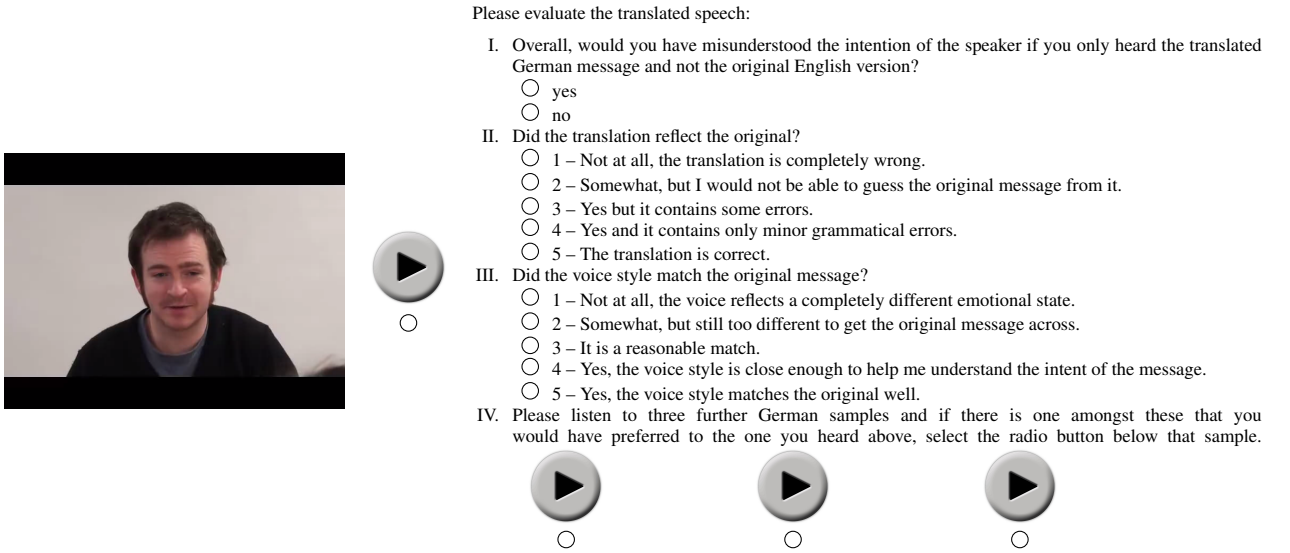

Fig. 8 Example of one trial from the FEAST online evaluation. The video utterance is shown on the left, the output of the affective speech translation is presented as a "play" button in the middle, and the full questionnaire is reproduced on the right.

II. on a 5-point Likert scale, how well the translation reflected the original utterance;
III. on a 5-point Likert scale, how well the synthetic voice style matched the original message;
IV. which one (if any) of the three available alternative styles the participant might have preferred.

Details of the trial layout and questionnaire are reproduced in Figure 8.

The evaluation was completed by 20 participants (8 f/12 m), 19 of which are native speakers of German.

In the answers to the first question, 4 out of the 20 sentences received the score that they would have been misunderstood without the presence of the English translation, by more than 25 % of the participants.[12] The overall score for translation was 3.98 on average, and for the suitability of the voice style 3.06. The average translation accuracy scores for the individual sentences, as well as the portion of sentences that were indicated to potentially have been misunderstood, is displayed in Figure 9. The scores for the frequently "misunderstood" sentences were 2.39 and 2.86, respectively. When looking at all individual cases of indicated possible misunderstandings, the translation receives scores of 2.19, and the voice style, 2.86 (as opposed to non-misunderstood sentences scoring 4.38 and 3.10, respectively). This indicates that both voice style and translation have an effect on the possibility of a sentence being misunderstood.

[12] Out of the 20 subjects, 3 seem to have misinterpreted the first question and answered the polar question consistently with the opposite value than intended. Because the data clearly showed the consistent reversal of *yes* and *no* responses to question I, their results were corrected accordingly.

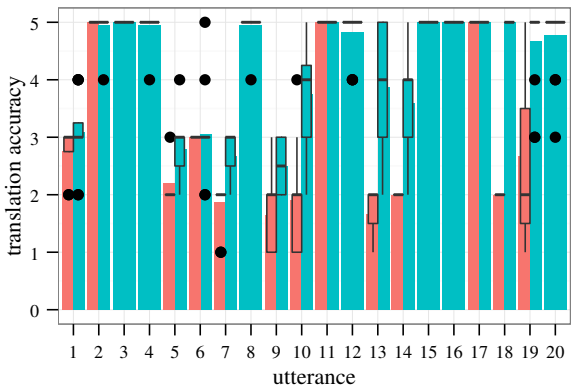

Fig. 9 Mean translation accuracy per utterance in the online evaluation as perceived by participants. Red bars indicate the response that the message would have been lost without reference to the original (cf. Figure 8, question I.); blue bars, that the message was preserved in the translation.

The answers to the fourth question reveal that listeners agreed with the voice selection of the FEAST system 52.3 % of the time. In all other cases, they indicated a preference for a different voice style. In the cases where the subjects agreed with the voice style selected by the system, only 14.5 % of the sentences were marked as misunderstood, as opposed to 21.6 % of misunderstood cases where the subject did not agree with the voice style. This result may indicate that in some cases, the selection of the correct voice style by the system may have prevented a sentence from being misunderstood by a listener.

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 the evaluations that (depending on the speaker) *cheerful/happy* can often be mistaken for *neutral*. However, from a usability perspective this is much more acceptable than systematic confusion of either with negative affect.

The evaluation with machine translation output indicated that the selection of a correct voice style does not only help capture the emotional intent of a message, but in cases where machine translation errors are present, it may aid the correct understanding of the sentence by the listener.

This paper has presented the FEAST system for affective speech-to-speech translation, which draws on user facial expression to incorporate appropriate expressiveness into the synthetic speech output in the target language.

Future additions to the full system include integration of prosodic features extracted from the acoustic input early in the processing pipeline to enhance the robustness of the affective state analysis. Furthermore, textual content analysis could also help to analyse the user's affective state for the target speech.

Acknowledgements 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 (http://semaine-db.eu/) [12].

References

1. Agüero P.D., Adell J., Bonafonte A.: Prosody generation for speech-to-speech translation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. I–557–560 (2006). doi:10.1109/ICASSP.2006.1660081
2. Ahmed Z., Steiner I., Székely É., Carson-Berndsen J.: A system for facial expression-based affective speech translation. In: ACM International Conference on Intelligent User Interfaces Companion, pp. 57–58 (2013). doi:10.1145/2451176.2451197
3. Batliner A., Huber R., Niemann H., Nöth E., Spilker J., Fischer K.: The recognition of emotion. In: W. Wahlster (ed.) Verbmobil: Foundations of Speech-to-Speech Translations, pp. 122–130. Springer (2000)
4. Breiman L., Friedman J.H., Olshen R.A., Stone C.J.: Classification and Regression Trees. Wadsworth (1984)
5. Chang C.C., Lin C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2**(3), 27:1–27:27 (2011). doi:10.1145/1961189.1961199
6. Cowie R., Douglas-Cowie E., Apolloni B., Taylor J.G., Romano A., Fellenz W.: What a neural net needs to know about emotion words. In: World Multiconference on Circuits, Systems, Communications and Computers, pp. 109–114 (1999)
7. Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., Taylor J.G.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine **18**(1), 32–80 (2001). doi:10.1109/79.911197
8. Ekman P., Keltner D.: Universal facial expressions of emotion: an old controversy and new findings. In: U. Segerstråle, P. Molnár (eds.) Nonverbal Communication: Where Nature Meets Culture, pp. 27–46. Lawrence Erlbaum (1997)
9. Kano T., Sakti S., Takamichi S., Neubig G., Toda T., Nakamura S.: A method for translation of paralinguistic information. In: International Workshop on Spoken Language Translation (2012)
10. Küblbeck C., Ernst A.: Face detection and tracking in video sequences using the modified census transformation. Image and Vision Computing **24**(6), 564–572 (2006). doi:10.1016/j.imavis.2005.08.005

11. Machado A.F., Queiroz M.: Techniques for crosslingual voice conversion. In: IEEE International Symposium on Multimedia (ISM), pp. 365–370 (2010). doi:10.1109/ISM.2010.62
12. McKeown G., Valstar M.F., Cowie R., Pantic M.: The SEMAINE corpus of emotionally coloured character interactions. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 1079–1084 (2010). doi:10.1109/ICME.2010.5583006
13. Och F.J.: Minimum error rate training in statistical machine translation. In: Annual Meeting of the Association for Computional Linguistics, pp. 160–167 (2003). doi:10.3115/1075096.1075117
14. Schröder M., Baggia P., Burkhardt F., Pelachaud C., Peter C., Zovato E.: EmotionML – an upcoming standard for representing emotions and related states. In: S. D'Mello, A. Graesser, B. Schuller, J.C. Martin (eds.) Affective Computing and Intelligent Interaction, pp. 316–325. Springer (2011). doi:10.1007/978-3-642-24600-5_35
15. Schröder M., Trouvain J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology **6**(4), 365–377 (2003). doi:10.1023/A:1025708916924
16. Shin J., Georgiou P.G., Narayanan S.: Enabling effective design of multimodal interfaces for speech-to-speech translation system: An empirical study of longitudinal user behaviors over time and user strategies for coping with errors. Computer Speech & Language **27**(2), 554–571 (2013). doi:10.1016/j.csl.2012.02.001
17. Steiner I., Schröder M., Charfuelan M., Klepp A.: Symbolic vs. acoustics-based style control for expressive unit selection. In: ISCA Tutorial and Research Workshop on Speech Synthesis (SSW), pp. 114–119 (2010)
18. 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: Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), pp. 5–8 (2012)
19. Tomás J., Canovas A., Lloret J., García M.: Speech translation statistical system using multimodal sources of knowledge. In: International Multi-Conference on Computing in the Global Information Technology (ICCGI), pp. 5–9 (2010). doi:10.1109/ICCGI.2010.26
20. Vauquois B.: Automatic translation – a survey of different approaches. In: S. Nirenburg, H.L. Somers, Y. Wilks (eds.) Readings in Machine Translation, chap. 28, pp. 333–338. MIT Press (2003)
21. Wöllmer M., Metallinou A., Eyben F., Schuller B., Narayanan S.: Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In: Interspeech, pp. 2362–2365 (2010)