PROGRESS IN FACIAL EXPRESSION BASED AFFECTIVE SPEECH TRANSLATION

Zeeshan Ahmed[1], *Ingmar Steiner*[2-3], *Éva Székely*[1], *and Julie Carson-Berndsen*[1]

[1]*Centre for Next-Generation Localisation, University College Dublin*
[2]*Multimodal Computing and Interaction, Saarland University*
[3]*DFKI GmbH, Saarbrücken*

zeeshan.ahmed@ucdconnect.ie

Abstract: Speech-to-speech translation is an emerging field for applications in spoken language technology. Translation systems currently focus on the processing of linguistic content, without taking into account the significance of paralinguistic information conveyed by visual gestures in human face-to-face communication. In the project presented in this paper, we have implemented a speech-to-speech translation system which preserves information about the user's affective state by transmitting it through the processing pipeline to the output component, which renders the translated content in the appropriate speaking style using expressive speech synthesis.

1 Introduction

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 [12], enhancing the user's experience with the system through a multimodal interface [8], 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 [5].

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, ADELL, and BONAFONTE [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. [3] 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 affective 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. Real-time facial expression analysis has been used in applications such as emotion classification and interpretation of affective and social signals on video data [13]. The results of automatic facial expression recognition have also been directly connected to expressive speech synthesis in a speech synthesis platform designed for people with speech impairments [10]. In that system, still images of the user's face were captured at the time of typing a message, and the facial expression analysis determined the expressive style of the speech synthesiser.

In our current work, we aim to preserve paralinguistic information of the input speech in a speech-to-speech translation process, by extending this approach. Hereby, real-time video of the user's face is analysed as he is talking into the microphone, and the results are mapped to a speaking style of a speech synthesiser in the target language. This advancement poses several additional challenges which will be discussed in detail in the remainder of this paper.

2 System architecture and processing workflow

The Facial Expression-based Affective Speech Translation (FEAST) 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 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 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 Previous work

In our previous FEAST prototype system [11], we presented an integration of a facial expression analysis component with a German speech synthesis voice capable of generating spoken output in a number of different expressive styles. The feasibility of the FEAST system was demonstrated using video data from the SEMAINE corpus [6], both in an automatic evaluation and a perceptual study using multimodal data.

However, the prototype system included only the paralinguistic processing components (cf. Figure 1), while speech recognition and translation were simulated with mock-up components. The paralinguistic processing unit is composed of the components shown at right of Figure 1. The following sections explain each of these components.

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

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

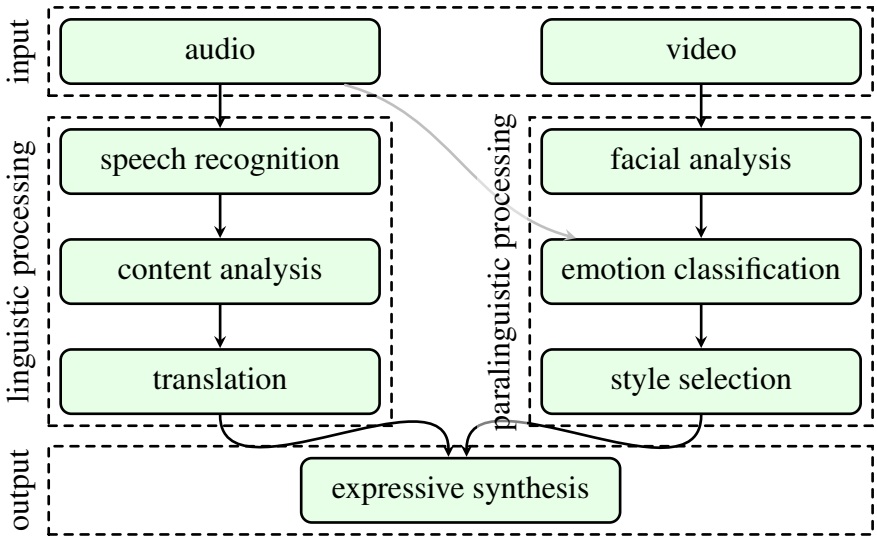

Figure 1 – System architecture of FEAST. The audio is not yet used for emotion classification.

3.1 Face detection and analysis

The face detection and expression analysis used in this study is performed by the Sophisticated Highspeed Object Recognition Engine (SHORE) library for real-time face detection and fine analysis.[3] 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 [4]. 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 [10], where static images were processed for utterance production. Because 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.[4]

3.2 Style selection

After the affective 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 category to voice style. Videos classified as *happy* are synthesised in a *cheerful* style, *sad* as *depressed*, and *angry* as *aggressive*, respectively. If the speaker's affective state is classified as *neutral*, the speech-to-speech translation results in a *neutral* speaking style.

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

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

3.3 Expressive speech synthesis

The text-to-speech (TTS) component uses the open-source synthesis platform MARY [7].[5] 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 [9]. A male German unit-selection voice which incorporates this feature is available;[6] 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 affective state analysis component is mapped onto one of the expressive styles available in the synthesis voice, which is added to the HTTP request as a parameter.

4 Advances

In the current work, we have implemented the speech-to-speech translation components (shown at left in Figure 1), using the Microsoft Speech SDK and Bing Translator API.

4.1 Linguistic processing components

In FEAST, the linguistic processing comprises speech recognition, content analysis, and machine translation components, as shown on the left 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 a MaryXML request.[7] Finally, this request is sent to the synthesis server, yielding the translation output, spoken in the target style.

4.2 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.

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

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

[7] http://mary.dfki.de/documentation/maryxml

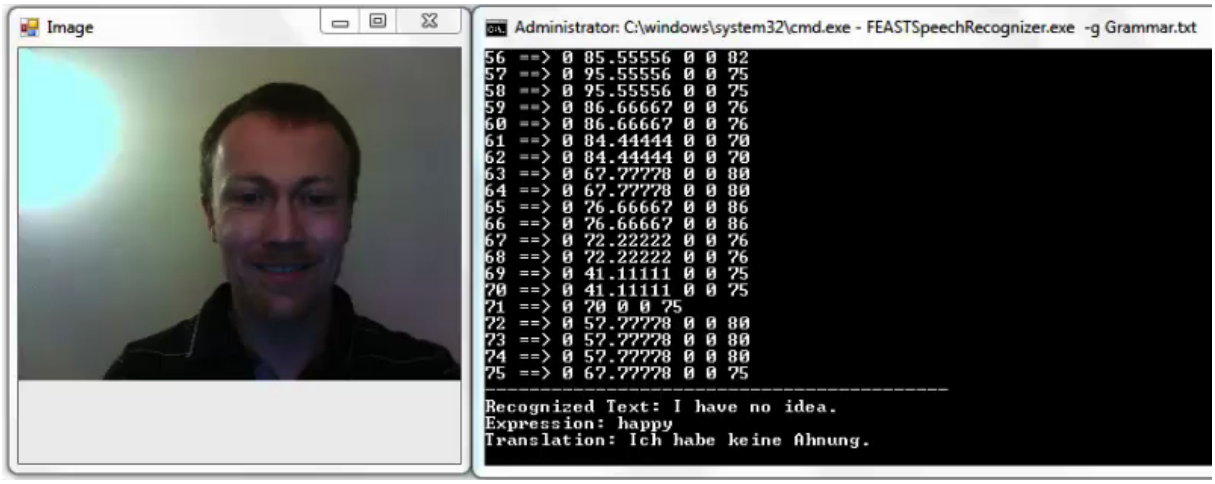

Figure 2 – Screenshot of the FEAST system in action (from [2]). The camera captures the user's face and displays it in the left window, while the console window to the right logs the recognised utterance and its translation, along with the facial expression classification results. The translation is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

5 Evaluation

An automatic evaluation of the FEAST system was performed on a subset of the SEMAINE corpus [6] using the annotation provided in the corpus. This automatic evaluation approach is complementary to the perceptual evaluation presented in [11], where human subjects were employed to rate the expression from the original video with the expressive synthetic speech-to-speech translation.

The purpose of the automatic evaluation is to highlight how accurately the system classifies the human emotion portrayed in the video. There are multiple dimensions for the evaluation of our system, e.g.,

- how accurate is the speech recognition component?
- how well does the translation component perform?
- what is the accuracy of the emotion classification?

However, our focus will only be on the accuracy of the facial expression classification, which is the core of the FEAST system. 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 classified. 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.

emotion	number of utterances	correct raw	correct 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.

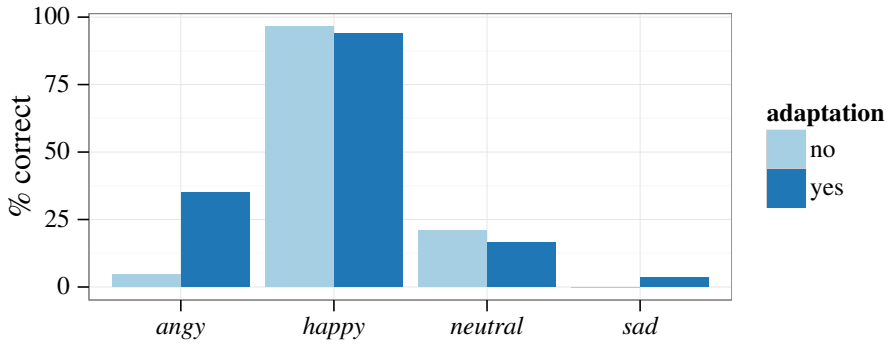

Figure 3 – Correct classification of utterances by portrayed emotion in SEMAINE corpus, before (light) and after (dark) adaptation.

5.1 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. 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 3.

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.

6 Conclusion and future work

This paper has presented the FEAST system for speech-to-speech translation, which draws on user facial expression to incorporate appropriate expressiveness into the synthetic speech output in the target language. The advances reported here include the integration of speech recognition and machine translation systems, using the Microsoft Speech SDK and Bing Translation API, respectively. These new developments will enable us to evaluate the FEAST system more thoroughly in an interactive experiment.

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.

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 (http://semaine-db.eu/) [6].

References

[1] P. D. AGÜERO, J. ADELL, and A. BONAFONTE: "*Prosody generation for speech-to-speech translation*". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1. 2006, pp. 557–560. DOI: 10.1109/ICASSP.2006.1660081.

[2] Z. AHMED, I. STEINER, É. SZÉKELY, and J. CARSON-BERNDSEN: "*A system for facial expression-based affective speech translation*". In: *ACM International Conference on Intelligent User Interfaces (IUI)*. Santa Monica, CA, USA, 2013, in press.

[3] T. KANO, S. SAKTI, S. TAKAMICHI, G. NEUBIG, T. TODA, and S. NAKAMURA: "*A method for translation of paralinguistic information*". In: *International Workshop on Spoken Language Translation (IWSLT)*. Hong Kong, China, 2012.

[4] C. KÜBLBECK and A. ERNST: "*Face detection and tracking in video sequences using the modified census transformation*". In: *Image and Vision Computing* 24.6 (2006), pp. 564–572. DOI: 10.1016/j.imavis.2005.08.005.

[5] A. F. MACHADO and M. QUEIROZ: "*Techniques for crosslingual voice conversion*". In: *IEEE International Symposium on Multimedia*. Taichung, Taiwan, 2010, pp. 365–370. DOI: 10.1109/ISM.2010.62.

[6] G. MCKEOWN, M. F. VALSTAR, R. COWIE, and M. PANTIC: "*The SEMAINE corpus of emotionally coloured character interactions*". In: *IEEE International Conference on Multimedia and Expo (ICME)*. Singapore, 2010, pp. 1079–1084. DOI: 10.1109/ICME.2010.5583006.

[7] M. SCHRÖDER and J. TROUVAIN: "*The German text-to-speech synthesis system MARY: a tool for research, development and teaching*". In: *International Journal of Speech Technology* 6.4 (2003), pp. 365–377. DOI: 10.1023/A:1025708916924.

[8] J. SHIN, P. G. GEORGIOU, and S. NARAYANAN: "*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*". In: *Computer Speech & Language* 27.2 (2013), pp. 554–571. DOI: 10.1016/j.csl.2012.02.001.

[9] I. STEINER, M. SCHRÖDER, M. CHARFUELAN, and A. KLEPP: "*Symbolic vs. acoustics-based style control for expressive unit selection*". In: *Seventh ISCA Workshop on Speech Synthesis (SSW)*. Kyoto, Japan, 2010, pp. 114–119.

[10] É. SZÉKELY, Z. AHMED, J. P. CABRAL, and J. CARSON-BERNDSEN: "*WinkTalk: a demonstration of a multimodal speech synthesis platform linking facial expressions to expressive synthetic voices*". In: *Third Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*. Montreal, Canada, 2012, pp. 5–8.

[11] É. SZÉKELY, Z. AHMED, I. STEINER, and J. CARSON-BERNDSEN: "*Facial expression as an input annotation modality for affective speech-to-speech translation*". In: *Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction (MA3)*. Santa Cruz, CA, USA, 2012.

[12] J. TOMÁS, A. CANOVAS, J. LLORET, and M. GARCÍA: "*Speech translation statistical system using multimodal sources of knowledge*". In: *Fifth International Multi-Conference on Computing in the Global Information Technology (ICCGI)*. Valencia, Spain, 2010, pp. 5–9. DOI: 10.1109/ICCGI.2010.26.

[13] J. WAGNER, F. LINGENFELSER, and E. ANDRÉ: "*The social signal interpretation framework (SSI) for real time signal processing and recognition*". In: *Interspeech*. Florence, Italy, 2011, pp. 3245–3248.