A System for Facial Expression-based Affective Speech Translation

Zeeshan Ahmed
Centre for Next Generation Localisation (CNGL), University College Dublin (UCD)
Belfield, Dublin 4, Ireland
zeeshan.ahmed@ucdconnect.ie

Ingmar Steiner
Multimodal Computing and Interaction,
Saarland University / DFKI
Campus C7.4, 66123
Saarbrücken, Germany
ingmar.steiner@dfki.de

Éva Székely
CNGL, UCD
eva.szekely@ucdconnect.ie

Julie Carson-Berndsen
CNGL, UCD
julie.berndsen@ucd.ie

ABSTRACT

In the emerging field of speech-to-speech translation, emphasis is currently placed on the linguistic content, while the significance of paralinguistic information conveyed by facial expression or tone of voice is typically neglected. We present a prototype system for multimodal speech-to-speech translation that is able to automatically recognize and translate spoken utterances from one language into another, with the output rendered by a speech synthesis system. The novelty of our system lies in the technique of generating the synthetic speech output in one of several expressive styles that is automatically determined using a camera to analyze the user's facial expression during speech.

Author Keywords

Emotion and Affective User Interface; Multi-modal interfaces; Speech I/O; Video Analysis

ACM Classification Keywords

H.5.2. **[Information Interfaces and Presentation (e.g. HCI)]:** User Interfaces; I.2.7. **[Artificial Intelligence]:** Natural Language Processing

INTRODUCTION

One of the emerging fields for applications of speech and language technology is speech-to-speech translation, which combines automatic speech recognition, machine translation, and text-to-speech (TTS) synthesis in a single application. Users of such a system can speak in one language and listen to a spoken translation in another language, greatly facilitating face-to-face communication between users who have no language in common.

Current development on speech-to-speech translation, has focused almost exclusively on the linguistic aspects, i.e., the textual content of the translated utterances. Human communication, on the other hand, also conveys an abundance of paralinguistic information through additional channels, such as tone of voice, facial expressions, and gestures, all of which go beyond the verbal content. These paralinguistic aspects have been largely neglected in speech-to-speech translation systems, even when multiple modalities (e.g., audio and video) are available.

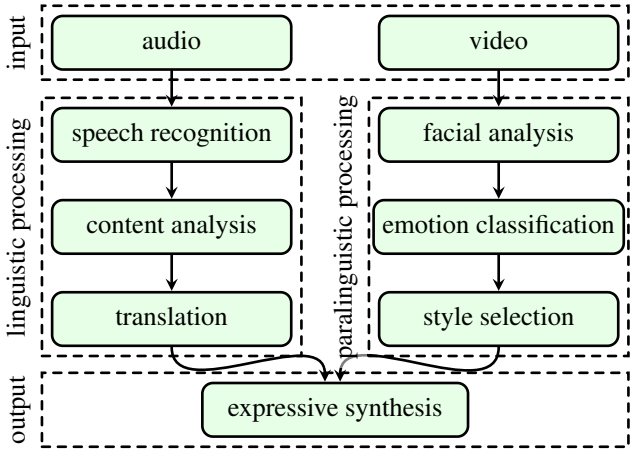

Figure 1. System architecture of FEAST

In previous work [6], we implemented a prototype system for Facial Expression-based Affective Speech Translation (FEAST), which is designed to analyze the facial expression of the user and synthesize the generated output in the target language using an expressive speaking style that corresponds to the user's affective state, e.g., for the benefit of another user who may not be able to see the former.

DEMONSTRATION

The demonstration of the FEAST system will consist of a short familiarization with the goals of affective speech translation followed by a hands-on experimentation with the system. Here, visitors will have the opportunity to speak any desired English sentence into the microphone, while their facial expressions are being analyzed; and listen to the German expressive synthetic speech output. A short video of the system in use will also be available for viewing.

IUI'13 Companion, March 19–22, 2013, Santa Monica, CA, USA.
ACM 978-1-4503-1966-9/13/03.

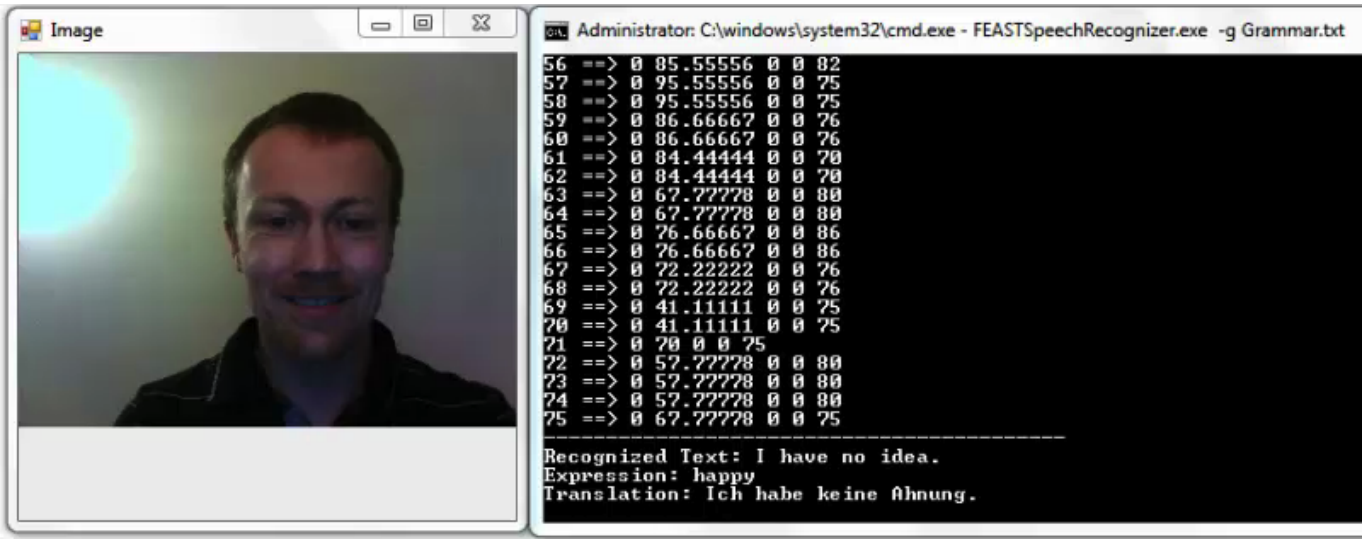

Figure 2. Screenshot of the FEAST system in action. The camera captures the user's face and displays it in the left window, while the console window to the right logs the recognized utterance and its translation, along with the facial expression classification results. The translation is synthesized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

An overview of the FEAST system's architecture is displayed in Figure 1, while a screenshot taken from a live demonstration of the system is shown in Figure 2.

Input

When running the FEAST system on a computing platform (i.e., a portable PC), the input is typically the live capture streams from the integrated camera and microphone. The audio and video streams are passed to the linguistic and paralinguistic components, respectively, for processing.

For testing purposes, it is also possible to provide video files as input, from which the audio and video streams are demuxed using FFmpeg [2] before they are processed as above.

Linguistic processing

Speech is recognized from the audio input using the Microsoft Speech software development kit (SDK) [3]. The Bing Translator application programming interface (API) [1] then translates the recognized text from the source language into the target language (in this demo, English to German).

Paralinguistic processing

The video stream is analyzed frame-by-frame using the Fraunhofer Sophisticated Highspeed Object Recognition Engine (SHORE) [4] to detect the user's face in the video frame. This library then classifies the facial expression of the user.

Expressive synthesis

The spoken output of the FEAST system is synthesized with the TTS platform MaryTTS [5], using a male German unit-selection synthesis voice specifically designed to support multiple expressive speaking styles.

The translated text generated by the linguistic processing components is wrapped in a MaryXML data structure, and the speaking style is selected by mapping the user's affective state, as determined by the SHORE classifier, to one of the available expressive styles *cheerful*, *depressed*, *aggressive*, or *neutral* (the default). The MaryXML input request is then transmitted to the MaryTTS server, which returns the expressive synthesis as an audio waveform. Finally, this synthesis output is played back to the user of the FEAST system. While these results are dependent on the capabilities of the speech synthesizer, it is equally possible to integrate another expressive speech synthesizer with different options and features.

CONCLUSION AND FUTURE WORK

The paralinguistic processing components in FEAST (Figure 1) have been evaluated using formal and subjective evaluation methods [6]. Extending the paralinguistic analysis with features extracted from the audio stream, as well as an extensive evaluation of the overall performance of the system, is the subject of future work.

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 and Trinity College Dublin.

REFERENCES

1. Bing Translator API. **http://www.microsofttranslator.com/dev/**.
2. FFmpeg multimedia framework. **http://ffmpeg.org/**.
3. Microsoft Speech SDK. **http://www.microsoft.com/en-us/download/details.aspx?id=10121**.
4. Küblbeck, C., and Ernst, A. Face detection and tracking in video sequences using the modified census transformation. *Image and Vision Computing 24*, 6 (2006), 564–572. **http://www.iis.fraunhofer.de/shore**.
5. Schröder, M., and Trouvain, J. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology 6*, 4 (2003), 365–377. **http://mary.dfki.de/**.
6. Éva Székely, Ahmed, Z., Steiner, I., and Carson-Berndsen, J. 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* (2012).