WinkTalk: a demonstration of a multimodal speech synthesis platform linking facial expressions to expressive synthetic voices

Éva Székely, Zeeshan Ahmed, João P. Cabral, Julie Carson-Berndsen
CNGL, School of Computer Science and Informatics, University College Dublin
Belfield, D4, Dublin, Ireland
{eva.szekely|zeeshan.ahmed}@ucdconnect.ie,{joao.cabral|julie.berndsen}@ucd.ie

Abstract

This paper describes a demonstration of the WinkTalk system, which is a speech synthesis platform using expressive synthetic voices. With the help of a webcamera and facial expression analysis, the system allows the user to control the expressive features of the synthetic speech for a particular utterance with their facial expressions. Based on a personalised mapping between three expressive synthetic voices and the users facial expressions, the system selects a voice that matches their face at the moment of sending a message. The WinkTalk system is an early research prototype that aims to demonstrate that facial expressions can be used as a more intuitive control over expressive speech synthesis than manual selection of voice types, thereby contributing to an improved communication experience for users of speech generating devices.

1 Introduction

During a human verbal communication process, expressive features of face and speech are congruent, operating in a synchronised manner (Campbell, 2008), (Graf et al., 2002). Facial expressions and expressive speech styles often help to convey the emotional intent of the speaker that is only partially contained in the words. The application described in this paper aims to make use of this synchrony and applies facial expressions as a real time volitional control over the expressive features of synthetic utterance productions of augmented speakers. The WinkTalk system is currently a research prototype in progress, operating on a personal computer equipped with a webcamera. The goal of the system is to respond to the need of integrated multimodality in speech generating devices of users of augmentative and alternative communication[1] (AAC) applications (Higginbotham, 2010). Being able to correctly link facial expression to synthetic speech output is a step forward to a more intuitive way of controlling the expressiveness of synthetic speech. The approach can be considered novel, as the authors are not aware of another system using facial expressions to control expressive TTS.

2 WinkTalk system architecture

The WinkTalk system is a web based application developed using AJAX and PHP technologies. The web application provides a flexible interface and allows for easy integration of new components such as synthetic voices or gesture recognisers running on a web server. The internal architecture of the system is shown in figure 1. The system operates based on a configurable workflow defining the three modes of the system: a personalisation mode, an automatic voice selection mode based on facial expression, which is the core functionality of the system, and a control mode of manual voice selection, that was included for evaluation purposes. In the manual voice selection application the user is presented with the three options and selects the voice style that

[1] Augmentative and alternative communication (AAC) refers to an area of research, clinical, and educational practice. AAC involves attempts to study and when necessary compensate for temporary or permanent impairments, activity limitations, and participation restrictions of individuals with severe disorders of speech-language production and/or comprehension, including spoken and written modes of communication.(ASHA, 2005)

NAACL-HLT 2012 Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), pages 5–8,
Montréal, Canada, June 7–8, 2012.

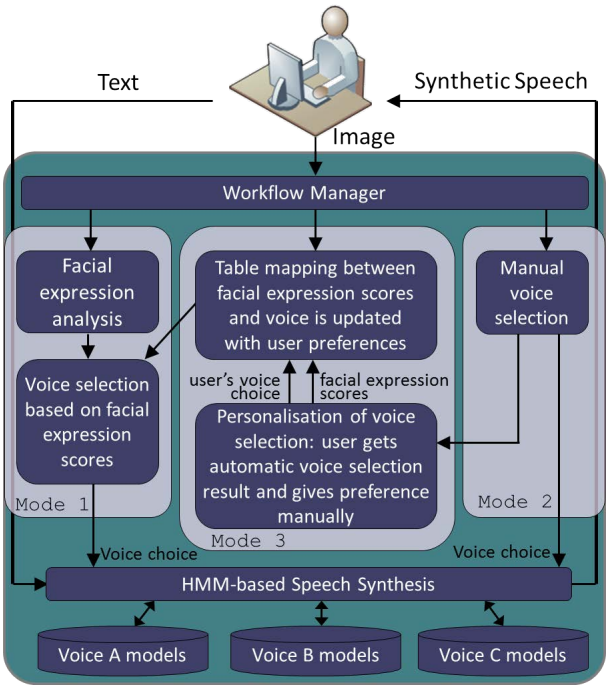

Figure 1: Architecture and working modes of the Wink-Talk system

matches the emotional or expressive intent of the message. It has previously been shown that after a short familiarisation with the voices, it is possible for the user to make a fairly good prediction of how a particular utterance will sound when synthesised with one of the voices (Székely et al., 2012). This makes it possible to use the system in a conversation situation, in which the user does not have the opportunity to listen to the three possible speech samples but needs to make a choice ahead of the time of the synthesis. The automatic voice choice mode and the personalisation mode will be described in sections 3 and 4, respectively.

3 Facial expression based voice selection

3.1 Expressive synthetic voices

The synthesiser component of the application uses three expressive HMM-based synthetic voices of a middle aged American male. The voices have been built using the HTS speech engine 2.1., from an audiobook corpus made available for Blizzard Challenge 2012 by Toshiba Research Europe Ltd, Cambridge Research Laboratory. Each synthetic voice was trained from different subcorpora of the audiobook obtained using an unsupervised clustering technique based on glottal source parameters (Székely et al., 2011). Perceptual experiments have shown (Hennig et al., 2012) that the three voices can be characterised on an expressiveness gradient: from calm (A voice), through intense (B voice) to very intense (C voice). This expressiveness gradient can be described with characteristics such as with rising pitch, greater prosodic variation, increased power and voice quality changing from lax to tense.

3.2 Facial expression analysis

For facial expression recognition, the system uses the Sophisticated Highspeed Object Recognition Engine (SHORE) library by Fraunhofer. To detect faces and expressions, SHORE analyses local structure features in images and outputs scores for four distinct facial expressions: *happy, sad, angry* and *surprised*, with an indication of the intensity of the expression (Kueblbeck and Ernst, 2006). The intensity ranges from 0-100, a higher value meaning a more intense expression in that category.

3.3 Mapping between facial expressions and voices

The system uses the facial expression categories and intensity scores outputted by SHORE to select from the three synthetic voices. The initial mapping between facial expression categories and ranges of intensity values and voices are shown in Table 1. For example, an image analysed as containing the facial expression *suprised* with an intensity of 25, the system will synthesise the corresponding utterance with the C voice. The system always uses the facial expression category with the highest value for a particular image. These initial values have been

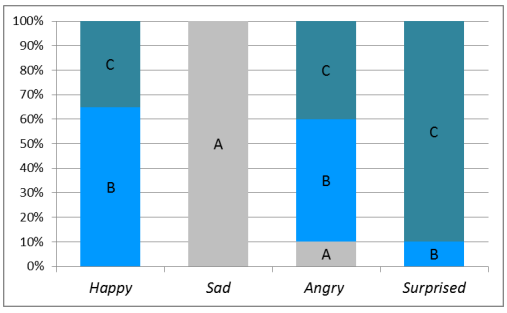

Figure 2: Initial thresholds for mapping different intensity values of the facial expressions to the synthetic voices

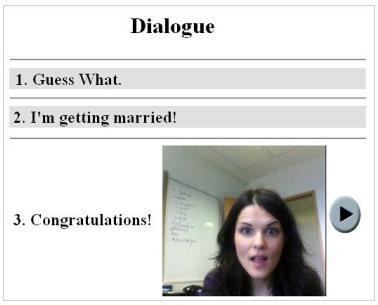

Figure 3: Interface of the dialogue simulation with WinkTalk.

chosen based on considerations about arousal levels of the underlying basic emotion of the facial expression categories, for example with surprise being a high arousal emotion, the intensity scores of it result sooner in a higher intensity voice choice. The values have also been supported by the results a perceptual test carried out by 25 participants on a dataset that was balanced to contain equal amount of stimuli from all facial expression categories. Participants were asked to select from three synthesised utterances the one best matching the facial expression of a person on a picture. The perceptual test has shown that 90% of all majority votings (above 66% agreement among participants) fell within the initial threshold values. When a message is being sent to the synthesiser, the system makes a snapshot of the user's face. Based on the image scores and threshold table, the system decides which voice best suits the current facial expression and returns the results accordingly. The system also provides an option to take streaming video input from the camera rather than a single image, and calculate the feature values over an interval of the video around the time of sending a message. To take into consideration the cases where individual preferences of voice choice differ greatly, as well as to account for individual differences in facial characteristics, a personalisation component has been integrated in the system, which will be introduced in section 4.

4 Personalisation component

In order to optimise the performance of the WinkTalk system, a personalisation session needs to be completed by each user. The objective of the personalisation is to adjust the voice selection threshold according to users' facial characteristics and individual preferences. In the personalisation phase,

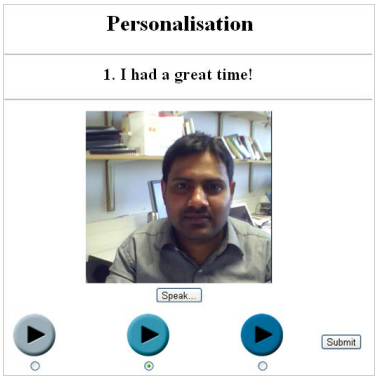

Figure 4: Interface of the personalisation component of WinkTalk.

the user is presented with a sentence and makes an appropriate facial expression to accompany the utterance. The facial expression is captured and analysed by the system and the user is presented with the three options of synthetic speech samples, along with an indication of which sample the system chose to match their facial expression. If the user does not agree with the selection provided by the system, a preference can be indicated by choosing from the other two options. The system then adjusts the threshold by moving it by a standard factor towards the outlying training example. The new threshold is applied in next trial. The thresholds for each facial expression-voice pair are normalised so that there is no overlap between the different voices for the same feature.

5 Conclusions and future work

The WinkTalk system has been evaluated within an interactive evaluation session involving 10 subjects, each of them acting out pre-scripted dialogues with a conversation partner. The evaluation has shown that while there is a general preference to manual selection of expressive voices, 90% of the participants described facial expression control as a valuable addition to the simulated augmented communication process. A strong learning effect in the ease of using the system has also been observed. Future work is planned to research further input strategies of gestures as well as to integrate a female expressive synthetic voice. An essential next step is to extend the

personalisation component to include the possibility of fully personalised training of the facial expression analysis to fit individual needs of users who are restricted with respect to their gestural expressiveness.

6 Demonstration

6.1 Overview

The demonstration will give participants an opportunity to use the WinkTalk system by conducting the personalisation phase and using the system with pre-scripted dialogues. It is intended for those interested in using multimodal tools and expressive speech to improve the communication experience of individuals with complex communication needs. The demonstration will give participants a chance to experience the facial expression control over the voice choice of the system as well as get an impression of how the range of expressive voices can be used in an acted dialogue situation. A 3 minute video of the system in use will also be available for viewing.

6.2 Familiarisation/Personalisation phase

First, a short introduction will be given to the system and its aims, then the participants will be introduced to the synthetic voices by listening to a few samples receiving a brief description of their characteristics. Subsequently, the participants will be asked to conduct a personalisation session including 20 iterations, that will help optimise the system to adapt to the participants' preferences, as described in section 4. It will also familiarise the users with the characteristics of the voices and the mapping of facial expressions and voices.

6.3 Dialogue simulation with synthetic voices

After the users are familiarised with the system, they can choose from a set of 8 dialogues representing a range of social interactions and emotional sentiment and intensity. Participants will act out some of the dialogues with a conversation partner, using facial expressions to control the selection of the synthetic voices instead of speaking with their own voice. They will also have the option to compare the facial expression control of the WinkTalk system with a simple manual selection of synthetic voices for each utterance. At the end of the dialogue session there will be a chance to fill out a feedback form to help the further development of the system.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin (UCD). 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. The authors would also like to thank Shannon Hennig (IIT), Nick Campbell and the Speech Communication Lab (TCD), for their invaluable help with the interactive evaluation.

References

American Speech-Language-Hearing Association. 2005 *Roles and Responsibilities of Speech-Language Pathologists With Respect to Augmentative and Alternative Communication: Position Statement.* Available from www.asha.org/policy.

Campbell, N. 2008. *Multimodal processing of discourse information; the effect of synchrony* Proc. of International Symposium on Universal Communication, Osaka.

Graf, H.P., Cosatto, E., Strom, V., and Huang, F.J. 2002. *Visual prosody: Facial movements accompanying speech.* Proc. of the 5th International Conference on Automatic Face and Gesture Recognition.

Hennig, S., Székely, É., Carson-Berndsen, J. and Chellali, R. 2012. *Listener evaluation of an expressiveness scale in speech synthesis for conversational phrases: implications for AAC.* to appear in: Proc. of ISAAC, Pittsburgh.

Higginbotham, D. J. 2010. *Humanizing Vox Artificialis: The Role of Speech Synthesis in Augmentative and Alternative Communication* Computer Synthesized Speech Technologies: Tools for Aiding Impairment, J. Mullennix and S. Stern, Eds. IGI Global, pp. 50-70.

HTS-2.1 toolkit, HMM-based speech synthesis system version 2.1. http://hts.sp.nitech.ac.jp.

Kueblbeck., C. and Ernst, A. 2006. *Face detection and tracking in video sequences using the modified census transformation.* Journal on Image and Vision Computing, vol. 24, issue 6, pp. 564-572.

SHORE face detection engine, Fraunhofer Institute http://www.iis.fraunhofer.de/en/bf/bsy/fue/isyst

Székely, É., Cabral, J., Abou-Zleikha, M., Cahill, P. and Carson-Berndsen, J. 2012. *Evaluating expressive speech synthesis from audiobook corpora for conversational phrases.* Proc. of LREC, Istanbul.

Székely, É., Cabral, J. P., Cahill, P. and Carson-Berndsen, J. 2011. *Clustering expressive speech styles in audiobooks using glottal source parameters.* Proc. of Interspeech, Florence.