AVSP 2009 - International Conference
on Audio-Visual Speech Processing
University of East Anglia, Norwich, UK
September 10-13, 2009

SynFace - Verbal and Non-verbal Face Animation from Audio

Jonas Beskow, Giampiero Salvi and Samer Al Moubayed

KTH Centre for Speech Technology, Stockholm, Sweden.
{beskow,giampi,sameram}@kth.se

Abstract

We give an overview of SynFace, a speech-driven face animation system originally developed for the needs of hard-of-hearing users of the telephone. For the 2009 LIPS challenge, SynFace includes not only articulatory motion but also non-verbal motion of gaze, eyebrows and head, triggered by detection of acoustic correlates of prominence and cues for interaction control. In perceptual evaluations, both verbal and non-verbal movmements have been found to have positive impact on word recognition scores.

1 Overview

SynFace (Beskow et al., 2004) employs a 3D animated face model controlled by articulatory oriented parameters. An online control model drives the system based on time-stamped phonetic input. We implemented a phonetic recognizer in order to drive the face movement from the incoming speech signal. The constraints on the recognizer for this application are speaker independence, task independence and, above all, low latency. The recognizer is based on a hybrid of hidden Markov models and recurrent neural networks. Special effort has been put in reducing all sources of latency in the processing chain. This was achieved by limiting the look-ahead in neural networks, decoder and the control model.

2 Non-verbal Cues

Recent work on SynFace has been aimed at improving the overall communicative experience through non-articulatory facial movements. We have chosen to focus on two classes of non-verbal movements that have been found to play important roles in communication and that also may be driven by acoustic features that can be reliably estimated from speech. The first category is speech-related movements linked to emphasis or prominence, the second category is gestures related to interaction control in a dialogue situation. SynFace uses an unsupervised approach to detecting prominence on a continuous scale using vowel acoustic features over time. By modeling vowel duration, mean loudness and mean delta F_0 over the vowel, a prominence level for each vowel is estimated using a function of the complement of the likelihood of these parameters, and the peaks of this function are hypothesized to estimate prominence of segments in their context. The model (mean and std) of these parameters is adapted in real-time to adjust to speaker change, speech rate, etc. For movements related to turn-taking, we are exploiting the fact that clause or turn ends often are associated with mutual gaze (Kendon, 1967), and that such regions often are correlate with low F_0 (Ward and Tsukahara, 2000).

3 Evaluation

For evaluation we have carried out perceptual experiments on human word recognition in noise, showing a significant intelligibility increase when using SynFace visual support, compared to the speech alone (Beskow et al., 2008). We also have recent findings indicating that prominence signaled via head nods or eyebrow movmements increase intelligibility further (Al Moubayed and Beskow, 2009). Finally, we have shown that people's turn-taking behaviour is roboustly affected by cues in the talking head.

References

Al Moubayed, S. and Beskow, J. (2009). Effects of visual prominence cues on speech intelligibility. In *submitted to AVSP 2009*.

Beskow, J., Granström, B., Nordqvist, P., Al Moubayed, S., Salvi, G., Herzke, T., and Schulz, A. (2008). Hearing at Home - communication support in home environments for hearing impaired persons. In *Proc. of Interspeech*, Brisbane, Australia.

Beskow, J., Karlsson, I., Kewley, J., and Salvi, G. (2004). SynFace - a talking head telephone for the hearing-impaired. In Miesenberger, K., Klaus, J., Zagler, W., and Burger, D., editors, *Computers Helping People with Special Needs*, pages 1178–1186. Springer-Verlag.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychol (Amst)*, 26(1):22–63.

Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207.