DEMONSTRATION OF AN HMM-BASED PHOTOREALISTIC EXPRESSIVE AUDIO-VISUAL SPEECH SYNTHESIS SYSTEM

Panagiotis Paraskevas Filntisis, Athanasios Katsamanis, Petros Maragos

School of ECE, National Technical University of Athens, 15773 Athens, Greece
Athena Research and Innovation Center, 15125 Maroussi, Greece
filby@central.ntua.gr, {nkatsam, maragos}@cs.ntua.gr

The usage of conversational agents is rapidly increasing in everyday life (cortana, siri, etc.). It has been shown that the inclusion of a talking face, increases the intelligibility of speech and the naturalness of human-computer interaction. Furthermore, an agent capable of expressing emotions has a stronger appeal to the human party and affects the interlocutor's emotional state.

The proposed demonstration is a Hidden Markov Model (HMM) based photorealistic audio-visual speech synthesis system, capable of expressing emotions [1, 2]. The system is capable of generating a talking head speaking in three emotions: happiness, anger, and sadness, plus in neutral speaking style. Further capabilities of the system include 1) the usage of HMM interpolation [3] in order to generate speech with mixtures of the original emotions (e.g., both anger and happiness), and speech with different levels of expressiveness (by mixing with the neutral emotion), 2) the usage of HMM adaptation [4], in order to adapt to a target emotion using only a few number of sentences.

Equipment In order to showcase our system we will use a laptop and speakers. The system will run fully on the laptop.

Demonstration Experience During the demonstration, viewers will have the opportunity to:

1. Watch videos of the talking head speaking in 3 different emotions (plus neutral) and see how the expressive talking head feels more natural compared to the talking head speaking in neutral style.
2. Watch the talking head speaking in two or more emotions at the same time, and see how the weights assigned to each emotion affects the outcome. It will also be of great interest to see which emotion each viewer perceives. In addition, through interpolation with the neutral emotion, viewers will be able to watch the talking head speak in different expressiveness levels for each emotion.
3. See how the neutral talking head can be adapted to speak in another emotion using only a few sentences, and how the number of sentences used affects the expressiveness of the resulting talking head.

1. REFERENCES

[1] P. P. Filntisis, A. Katsamanis, and P. Maragos, "Photorealistic Adaptation and Interpolation of Facial Expressions using HMMs and AAMs for Audio-Visual Speech Synthesis," in *Proc. ICIP*, 2017.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0.," in *Proc. ISCA SSW6*, 2007.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system," *Acoustical Science and Technology*, vol. 21, no. 4, 2001.

[4] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, Language Processing*, vol. 17, 2009.