Towards High Quality Video-Realistic Expressive Audio-Visual Speech Synthesis

Panagiotis Paraskevas Filntisis, Athanasios Katsamanis, Pirros Tsiakoulis, Petros Maragos

PII: S0167-6393(17)30041-9
DOI: 10.1016/j.specom.2017.08.011
Reference: SPECOM 2488

To appear in: *Speech Communication*

Received date: 22 January 2017
Revised date: 5 June 2017
Accepted date: 28 August 2017

Please cite this article as: Panagiotis Paraskevas Filntisis, Athanasios Katsamanis, Pirros Tsiakoulis, Petros Maragos, Towards High Quality Video-Realistic Expressive Audio-Visual Speech Synthesis, *Speech Communication* (2017), doi: 10.1016/j.specom.2017.08.011

Towards High Quality Video-Realistic Expressive Audio-Visual Speech Synthesis

Panagiotis Paraskevas Filntisis[a,b,*], Athanasios Katsamanis[a,b], Pirros Tsiakoulis[c], Petros Maragos[a,b]

[a]*School of Electrical and Computer Engineering, National Technical University of Athens, Zografou Campus, Athens 15773, Greece*
[b]*Athena Research and Innovation Center, Maroussi 15125, Greece*
[c]*INNOETICS LTD, Athens, Greece*

Abstract

High quality expressive speech synthesis has been a long-standing goal towards natural human-computer interaction. Generating a talking head which is both realistic and expressive appears to be a considerable challenge, due to both the high complexity in the acoustic and visual streams and the large non-discrete number of emotional states we would like the talking head to be able to express. In order to cover all the desired emotions, a significant amount of data is required, which poses an additional time-consuming data collection challenge. In this paper we attempt to address the aforementioned problems in an audio-visual context. Towards this goal, we propose two deep neural network (DNN) architectures for Video-realistic Expressive Audio-Visual Text-To-Speech synthesis (EAVTTS) and evaluate them by comparing them directly both to traditional hidden Markov model (HMM) based EAVTTS, as well as a concatenative unit selection EAVTTS approach, both on the realism and the expressiveness of the generated talking head. Next, we investigate adaptation and interpolation techniques to address the problem of covering the large emotional space. We use HMM interpolation in order to generate different levels of intensity for an emotion, as well as investigate whether it is possible to generate speech with intermediate speaking styles between two emotions. In addition, we employ HMM adaptation to adapt an HMM based system to another emotion using only a limited amount of adaptation data from the target emotion. We performed an extensive experimental evaluation on a medium sized audio-visual corpus covering three emotions, namely anger, sadness and happiness, as well as neutral reading style. Our results show that DNN-based models outperform HMMs and unit selection on both the realism and expressiveness of the generated talking heads, while in terms of adaptation we can successfully adapt an audio-visual HMM set trained on a neutral speaking style database to a target emotion. Finally, we show that HMM interpolation can indeed generate different levels of intensity for EAVTTS by interpolating an emotion with the neutral reading style, as well as in some cases, generate audio-visual speech with intermediate expressions between two emotions.

Keywords: audio-visual speech synthesis, expressive, hidden Markov models, deep neural networks, interpolation, adaptation

1. Introduction

Human-computer interaction (HCI) has exploded in the latest decades and, virtual or physical, intelligent agents[1] have become an integral part of a person's everyday life. These agents take part in a variety of applications, either of small significance such as everyday human tasks, or of great significance such as teaching assistants for pedagogical purposes (Johnson et al., 2000). Artificial Intelligence (AI) aims to maximize the naturalness of human-computer interactions so that the human forgets that he is interacting with a computer. Taking into account the fact that the main mode of human communication is speech, we understand that speech synthesis constitutes a vital part of AI for human-computer communication. Speech synthesis does not include only the generation of a human-like voice; speech is multimodal in nature (McGurk and MacDonald, 1976; Ekman, 1984) and important information is included in the visual stream of information (i.e., the human face, and more generally the human head and its movements) along with the acoustic stream. It has been shown that the inclusion of a visual stream of information increases the intelligibility of speech, especially under noise, even when the face is a virtual talking head (Sumby and Pollack, 1954; Ouni et al., 2007).

Achieving a high degree of naturalness in HCI is highly correlated with the ability of the agent to express emotions. Agents capable of expressiveness are more believable and life-like, thus have a stronger appeal to their interlocutor (Bates et al., 1994). In addition, expressive behavior itself contains important information (Ambady and Rosenthal, 1992) and affects the emotional state of the other party (Hatfield et al., 1993; Keltner and Haidt, 1999) and, consequently, its decision making (Schwarz, 2000; Bechara, 2004).

Strongly correlated with speech, emotion is conveyed multimodally (Darwin, 1871; Richard J. Davidson, 2002), so the agent must be capable of expressing emotion multimodally as well. It is also reasonable to assume that emotions should be

*Corresponding author
Email addresses: filby@central.ntua.gr (Panagiotis Paraskevas Filntisis), nkatsam@cs.ntua.gr (Athanasios Katsamanis), ptsiak@innoetics.com (Pirros Tsiakoulis), maragos@cs.ntua.gr (Petros Maragos)

[1]An agent is anything that can be viewed as perceiving its environment through sensors, and acting upon that environment through effectors (Russell and Norvig, 1995). Example of a physical agent is a robot, while of a virtual agent is an avatar.