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

Panagiotis P. Filntisis, Athanasios Katsamanis and Petros Maragos

School of ECE, National Technical University of Athens, 15773 Athens, Greece
Athena Research and Innovation Center, 15125 Maroussi, Greece

DEMONSTRATION BACKGROUND

- The usage of conversational agents is rapidly increasing in everyday life (cortana, siri, etc.) and ***speech synthesis (both acoustic & audio-visual)*** constitutes a vital asset for 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
- **We demonstrate an HMM based photorealistic audio-visual speech synthesis system, capable of:**
  - **Generation of audio-visual speech in three emotions: happiness, anger, and sadness, plus in neutral speaking, for the Greek language**
  - **Usage of HMM adaptation, in order to adapt to a target emotion using only a few number of sentences**
  - **Usage of HMM interpolation in order to generate speech with mixtures of the original emotions and speech with different levels of expressiveness (by mixing with the "neutral" emotion)**

HMM-BASED AUDIO-VISUAL SPEECH SYNTHESIS

TRAINING

- Extract acoustic - visual features
- Train HMMs with EM algorithm
- Cluster similar phonetic contexts

SYNTHESIS

- Analyze input text
- Generate features from HMMs
- Reconstruct audio and video

ACTIVE APPEARANCE MODELS (AAM)

The face of the agent is modeled by Active Appearance Models:

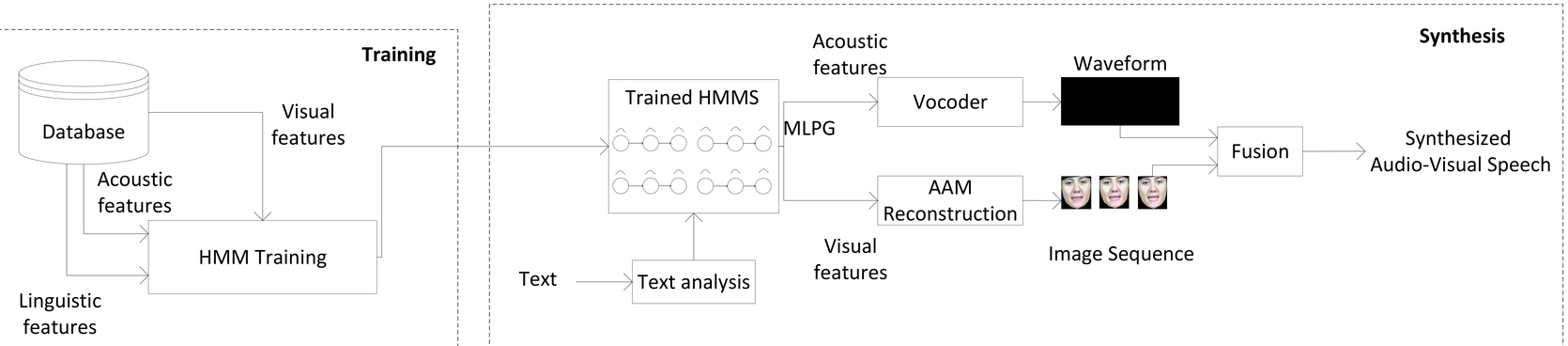
Face shape $\boldsymbol{s} = \bar{\boldsymbol{s}} + \sum_{i=1}^{n} p_i \boldsymbol{s_i}$

Face texture $\boldsymbol{A(x)} = \overline{\boldsymbol{A(x)}} + \sum_{i=1}^{l} \lambda_i \boldsymbol{A_i}$

$\bar{\boldsymbol{s}}$: mean shape
p_i: eigenshape coefficients
$\overline{\mathbf{A}(\boldsymbol{x})}$: mean texture
λ_i: eigentexture coefficients

DEMONSTRATION EXPERIENCE

- 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
- 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.
- 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.

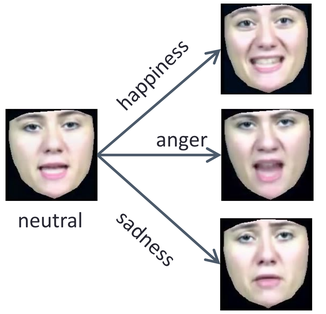


ADAPTATION

CSMAPLR adaptation is employed to adapt a neutral HMM system to another emotion using a small amount of adaptation sentences:

$$\bar{\boldsymbol{\mu}} = \boldsymbol{Z}\,\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \qquad \bar{\boldsymbol{\Sigma}} = \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^{\mathbf{T}}$$

$\boldsymbol{\mu}, \boldsymbol{\Sigma}$: original mean and covariance matrix
$\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}$: adapted mean and covariance matrix
$\boldsymbol{\varepsilon}, \mathbf{Z}$: transformation bias and matrix



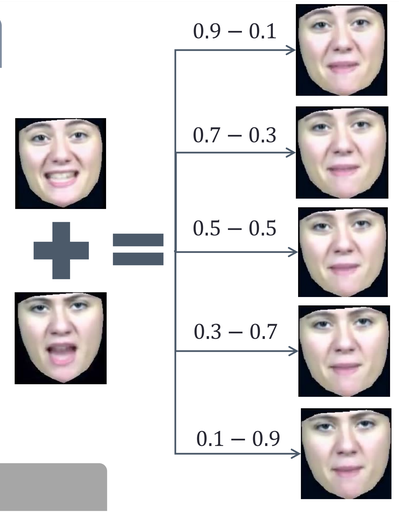
Adapting the neutral system to other emotions.

INTERPOLATION

Interpolation between observations is employed to interpolate statistics of HMMs from different HMM sets:

$$\boldsymbol{\mu} = \sum_{i=1}^{K} \alpha_i \boldsymbol{\mu_i}\,, \quad \boldsymbol{\Sigma} = \sum_{i=1}^{K} \alpha_i^2 \boldsymbol{\Sigma_i}$$

$\boldsymbol{\mu}, \boldsymbol{\Sigma}$: interpolated mean – covariance matrix
$\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}$: adapted mean – covariance matrix of *ith* HMM set
a_i: interpolation weight for *ith* HMM set



REFERENCES

[1] H. Zen et al., "The hmm-based speech synthesis system (hts) version 2.0.," in Proc. ISCA SSW6, pp. 294–299, 2007.

[2] J. Yamagishi et al., "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm,"IEEE Trans. Audio, Speech, Language Processing, vol. 17, pp. 66–83, 2009.

[3] T. Yoshimura et al., "Speaker interpolation for hmm-based speech synthesis system," Acoustical Science and Technology, vol. 21, pp. 199–206, 2001.

[4] P.P. Filntisis et al., "Video-realistic expressive audio-visual speech synthesis for the Greek language", 2017, https://doi.org/10.1016/j.specom.2017.08.011

This work has been funded by the BabyRobot project, supported by the EU Horizon 2020 Programme under grant 687831.