PHOTOREALISTIC ADAPTATION AND INTERPOLATION OF FACIAL EXPRESSIONS USING HMMS AND AAMS FOR AUDIO-VISUAL SPEECH SYNTHESIS

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

BACKGROUND & OBJECTIVES

- Intelligent agents have a continuous presence in everyday life and speech synthesis (both acoustic & audio-visual) constitutes a vital asset for human – computer interaction
- Achieving a high degree of naturalness in HCI depends on the ability of the agent to express emotions
- However, there is a huge data overhead when considering synthesis of expressive speech in a large non-discrete emotional space
- We tackle the problem this problem by:
  - ➢using HMM adaptation to adapt an existing audio-visual speech synthesis HMM set to a new emotion using a small amount of adaptation data
  - ➢employing HMM interpolation to combine HMM sets to generate speech with intermediate styles

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

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

mean texture | #1 eigentexture | mean + 3 sd eigentexture #1 | mean - 3 sd eigentexture #1

Example of the first eigentexture and the variations it causes to the mean texture

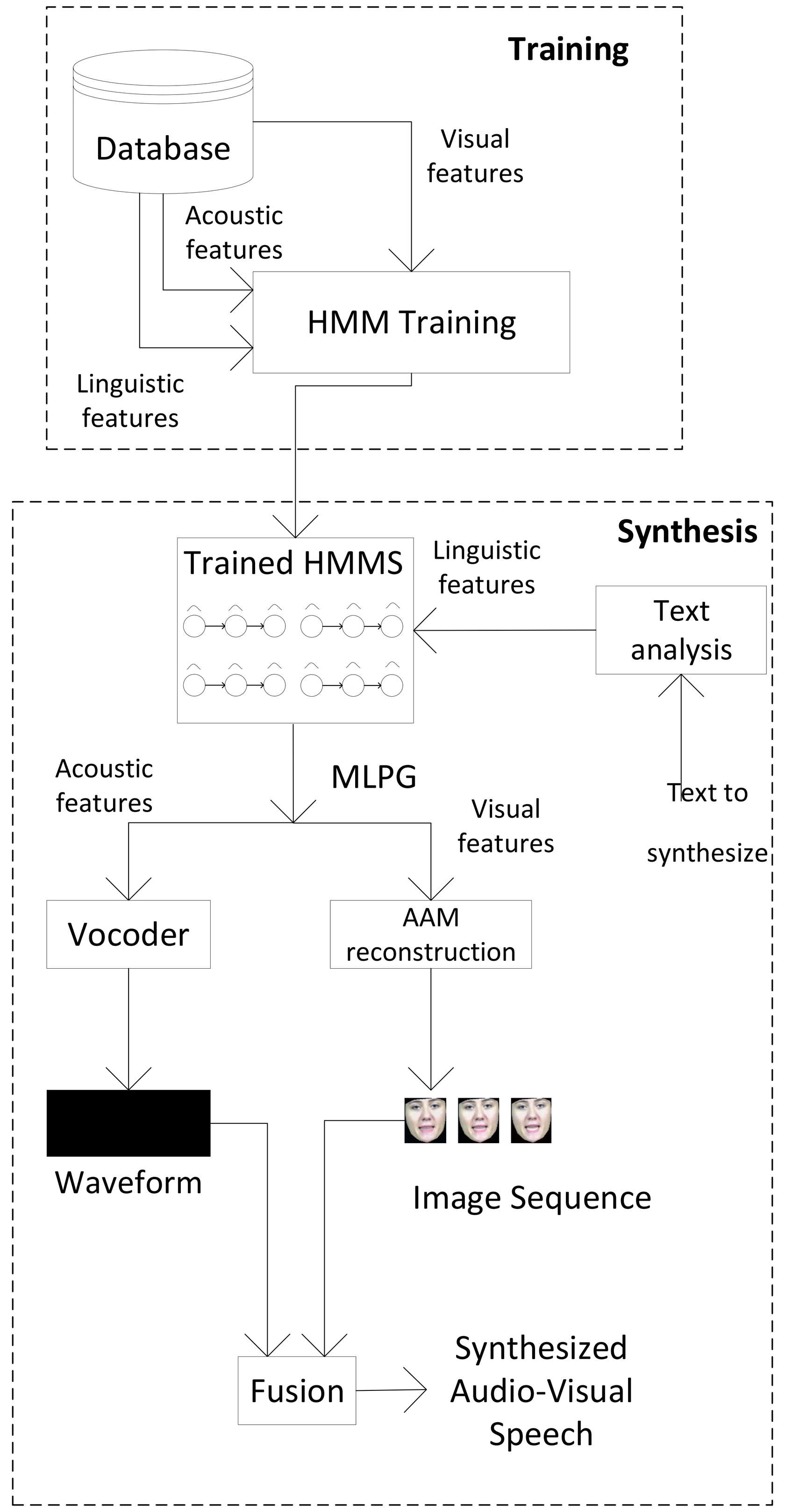
HMM-BASED AUDIO-VISUAL SPEECH SYNTHESIS [1]

TRAINING

- Extract acoustic and visual features
- Train HMMs with EM algorithm
- Cluster similar phonetic contexts using decision trees

SYNTHESIS

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



Adaptation

CSMAPLR ([2]) adaptation is employed to adapt a neutral emotion 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

Interpolation

Interpolation between observations ([3]) is employed to interpolate statistics of HMMs from different HMM sets:

$$\boldsymbol{\mu} = \sum_{i=1}^{K} \alpha_i \boldsymbol{\mu}_i\,, \qquad \boldsymbol{\Sigma} = \sum_{i=1}^{K} \alpha_i^2 \boldsymbol{\Sigma}_i$$

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

EXPERIMENTS & RESULTS

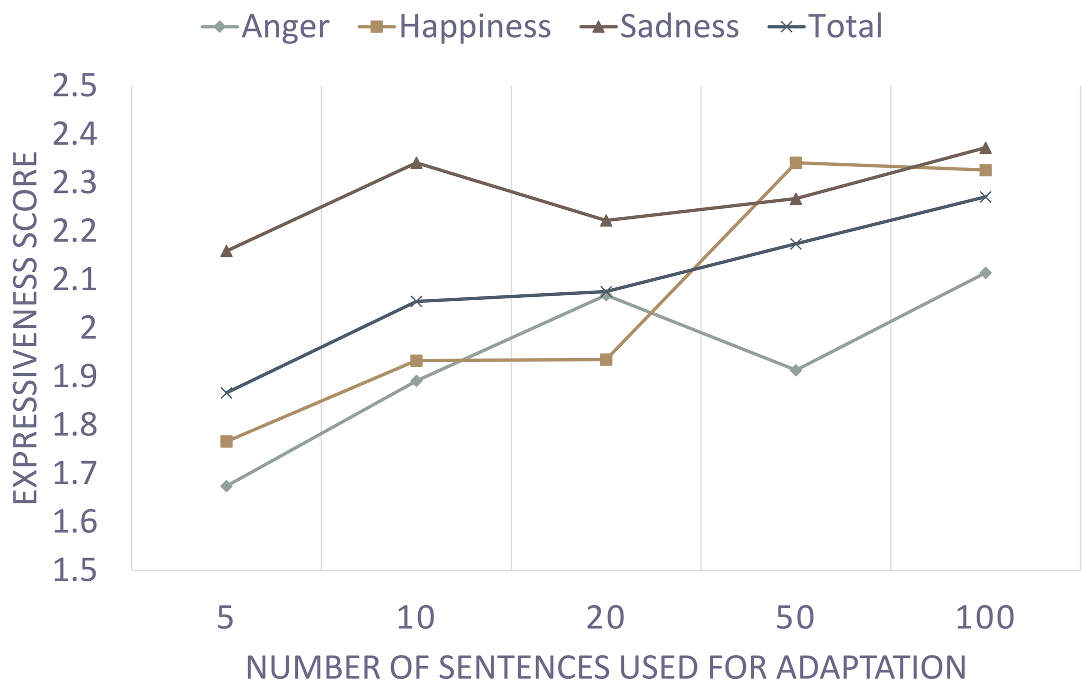
We trained four HMM-based audio-visual speech synthesis systems using the CVSP-EAV([4]) corpus which includes: *happiness, sadness, anger, neutral*.

First Evaluation

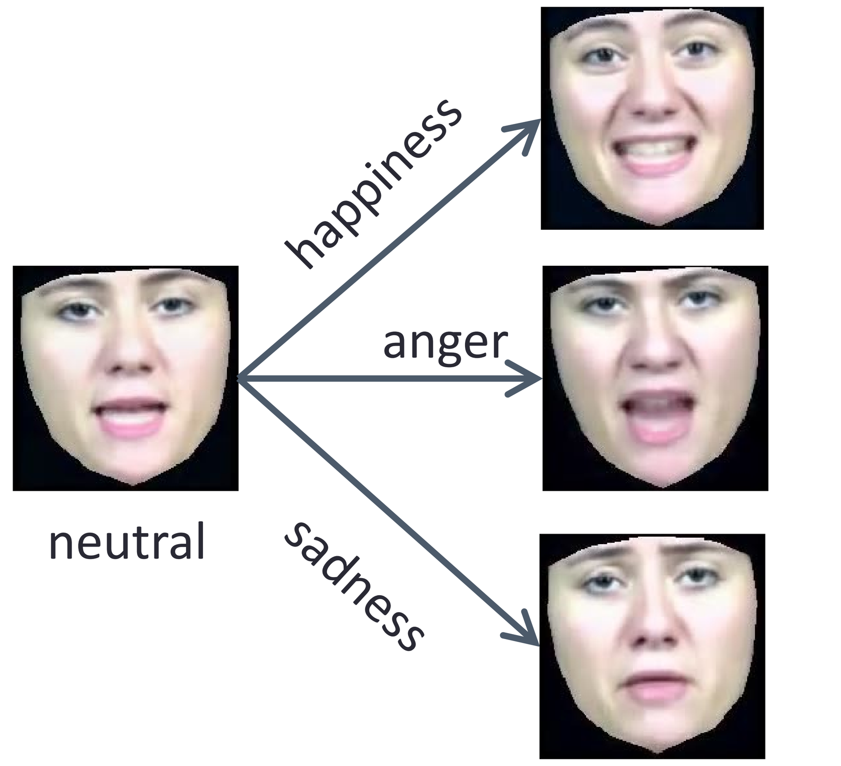
- We adapted the neutral HMM system to the other 3 emotions using a variable number of adaptation sentences.
- 32 humans evaluated the expressiveness of the agent on a discrete scale of 1 to 3 (increasing).

Second Evaluation

- We Interpolated the 6 emotion combinations for variable weight pairs: $(0.9, 0.1), (0.7, 0.3), (0.5, 0.5), (0.3, 0.7), (0.1, 0.9)$.
- 28 humans were asked to recognize the emotion in each combination/pair.



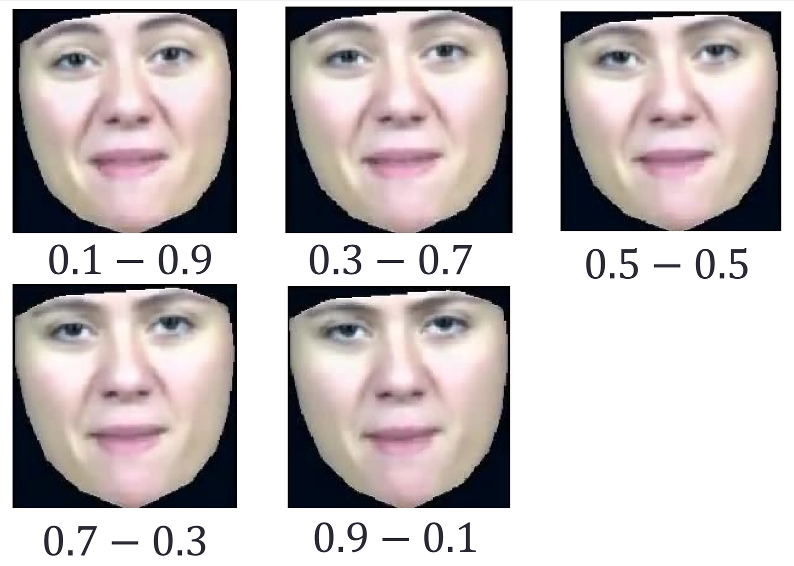
Subjective expressiveness score for a variable number of adaptation sentences.



Adapting the neutral emotion to other emotions.

Weights	Neutral	Anger	Happiness	Sadness
0.1 – 0.9	8.7	4.35	0	86.96
0.3 – 0.7	18.18	0	0	81.82
0.5 – 0.5	45.83	0	4.17	50
0.7 – 0.3	83.33	0	0	16.67
0.9 – 0.1	91.3	4.35	0	4.35

*Emotion classification rate when interpolating the **neutral** and **sadness** HMM systems (% scores).*



*Interpolating the **anger** and **happiness** HMM sets. (respective weights shown under each image).*

CONCLUSIONS

- We can successfully adapt an HMM-based audio-visual speech synthesis system to a target emotion using a small number of adaptation data. Level of expressiveness increases with number of adaptation sentences used.
- HMM interpolation gives us audio-visual speech with intermediate characteristics between the interpolated emotions.
- DNN version of the system can be found in [4].

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.