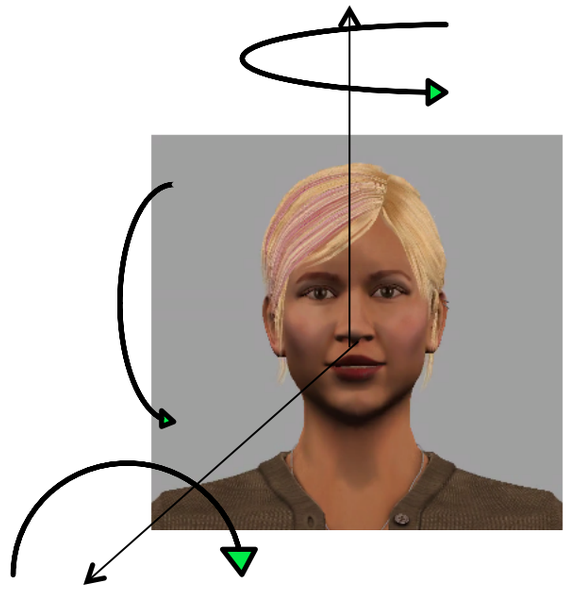
Head Motion Generation with Synthetic Speech: a Data Driven Approach

NAJMEH SADOUGHI AND CARLOS BUSSO

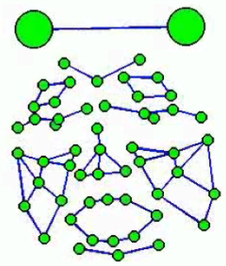
Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science



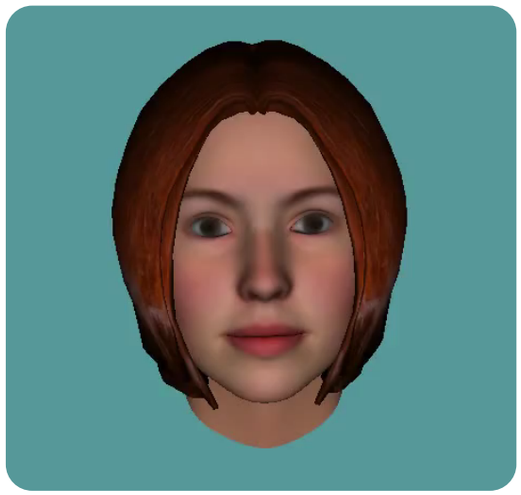
Sep, 2016

Motivation

- Head motion and speech prosodic patterns are strongly coupled
- Believable conversational agents should capture this relationship
  - Speech intelligibility [K. G. Munhall et al., 2004]
  - Naturalness [C. Busso et al., 2007, C. Liu et al., 2012, Marioryad et al., 2013]

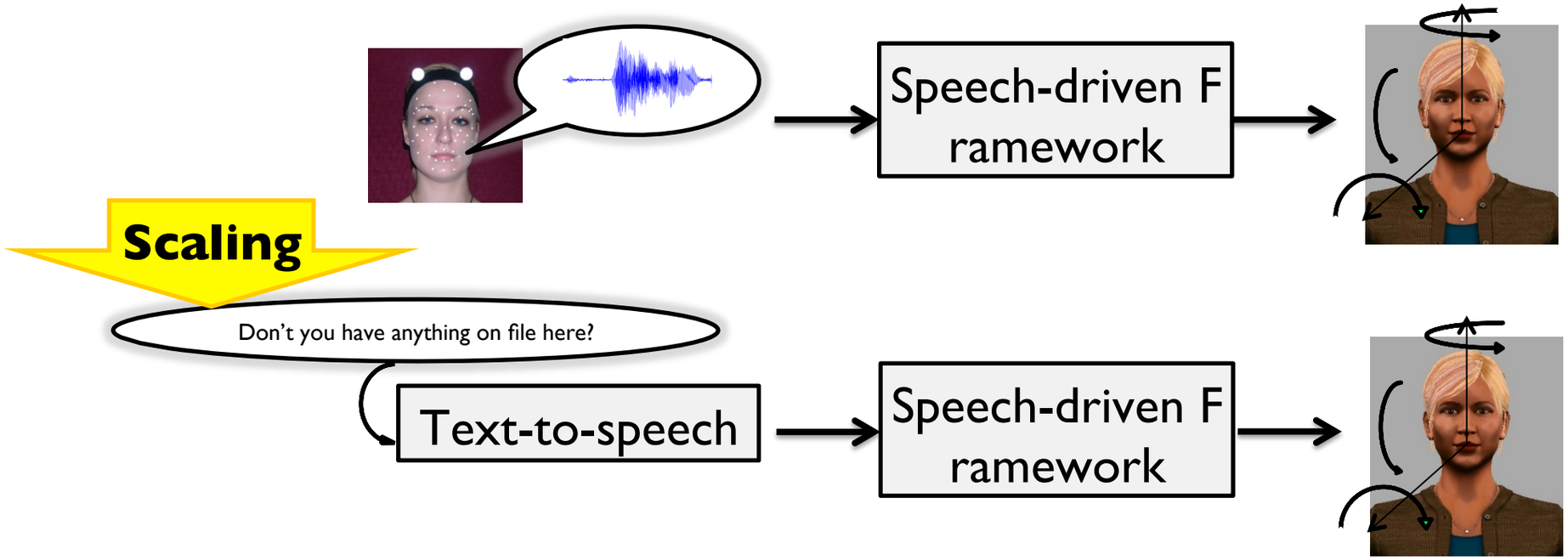


- Rule-based approaches
  - Rely on the content of the message to choose the movement
  - Synchronization with speech is challenging
- Speech-driven approaches
  - Learn the coupling from synchronized motion capture and audio recordings

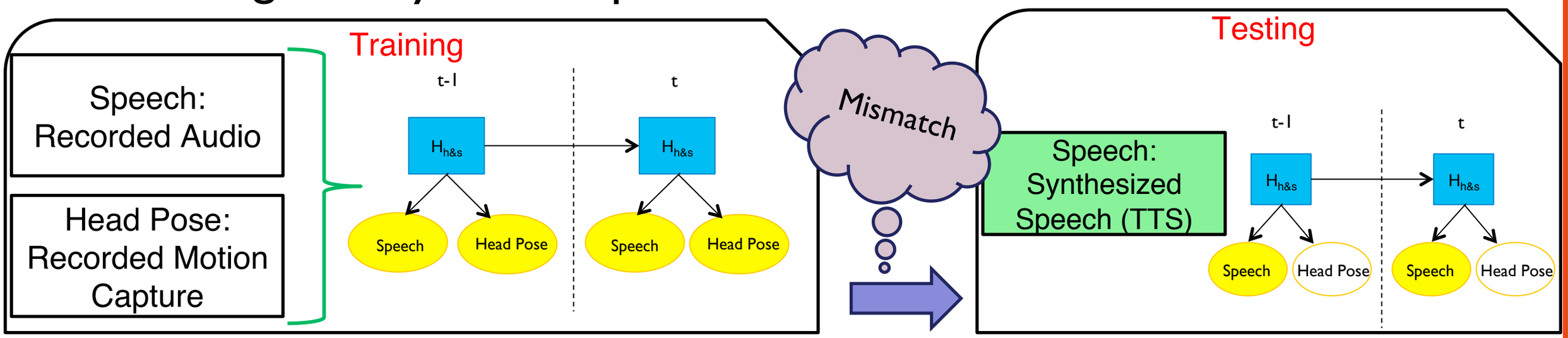


[Sadoughi et al., 2014]

Motivation

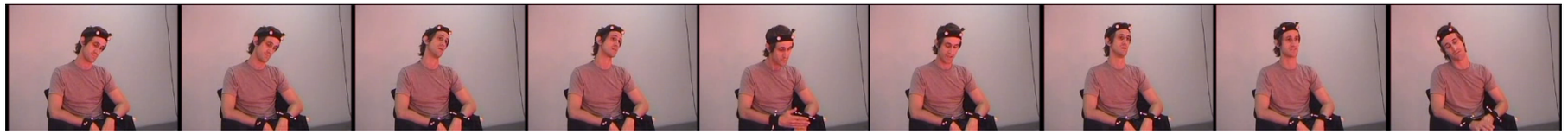


- Training with synchronized speech and head movement recording, testing with synthetic speech [Van Welbergen, Herwin et al., 2015]



- This paper addresses the problem with the mismatch

Overview

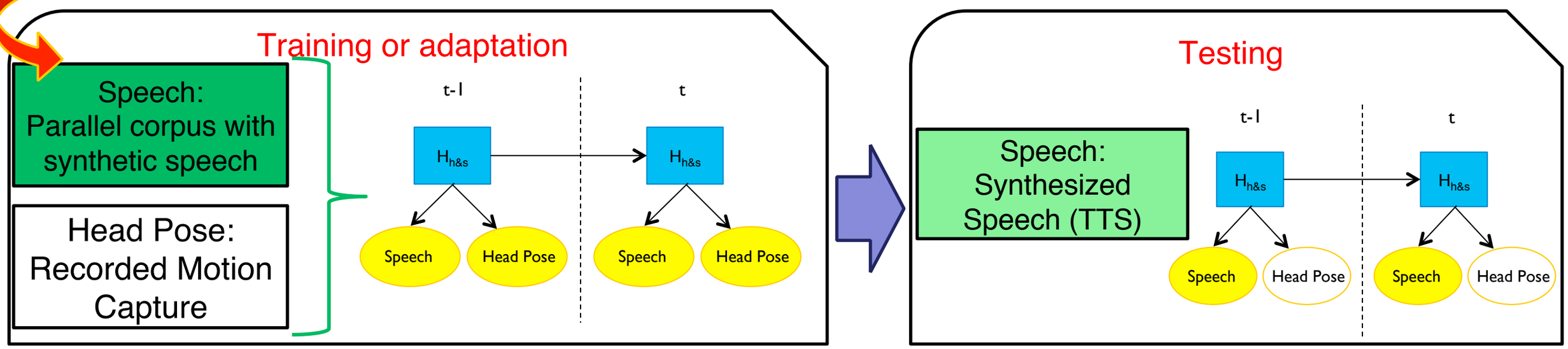


Original

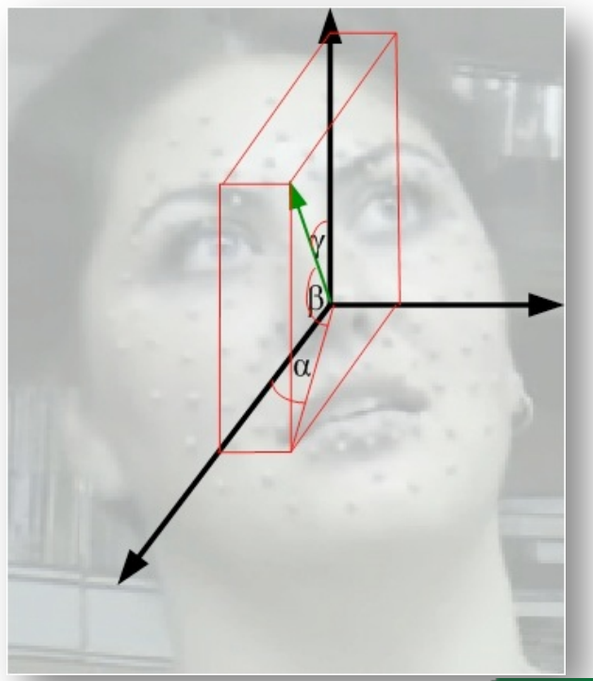
Synthesized

Our Proposal

Aligned



Corpus: IEMOCAP



- Video, audio and MoCap recording
- Dyadic interactions
- Script and improvisation scenarios
- We used 270.16 mins (non-overlapping speech)
- Three head angular rotations
- F0 and intensity (Praat)
- Mean normalization per subject
- Variance normalization, globally

Parallel Corpus

- OpenMary: open source text-to-speech (TTS)

Synthetic speech

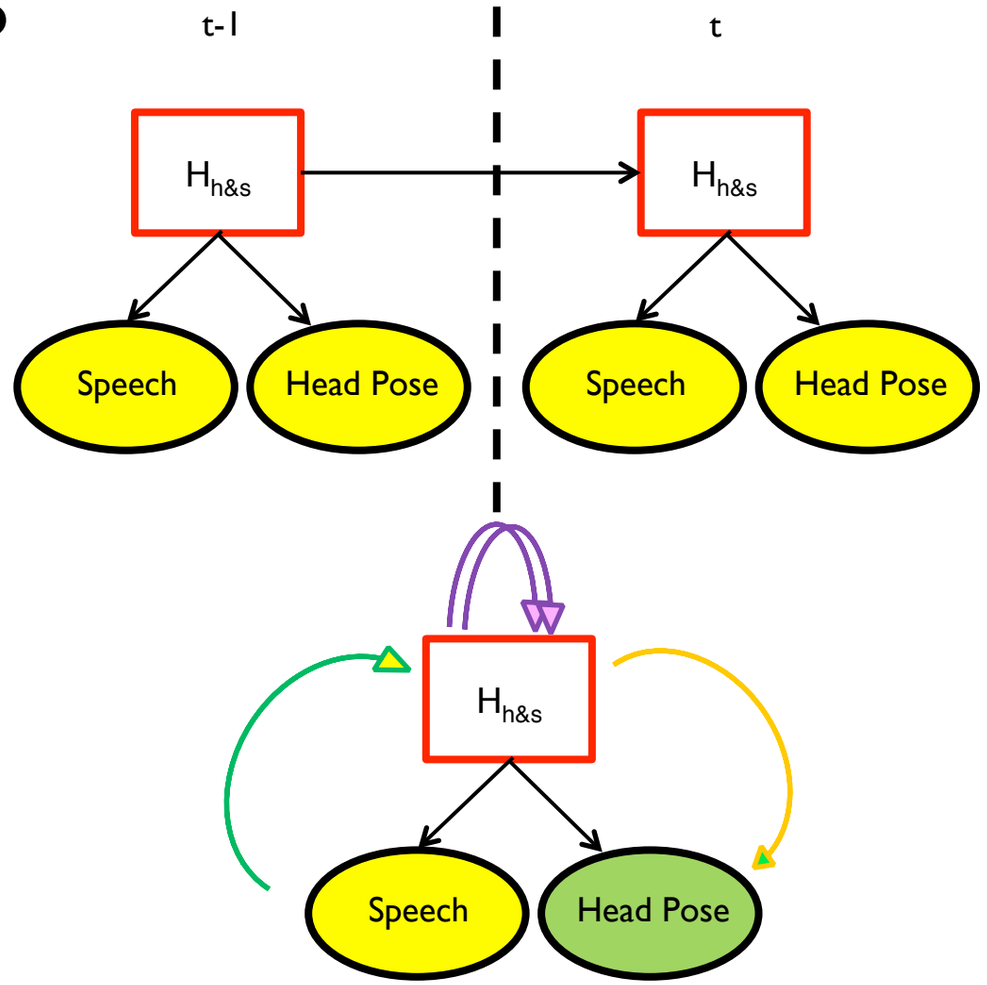
- Aligning the synthesized and original speech (word-level) [Lotfian and Busso, 2015]
- Praat warps the speech (pitch synchronous overlap add)
- Replacing the zero segments with silent recordings
- Mean normalization per voice
- Variance normalization to match the variance of the neutral segments in IEMOCAP

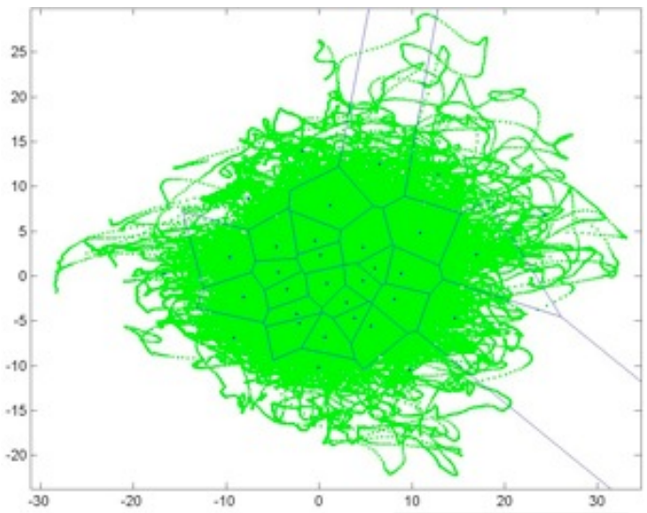
Original speech

Aligned synthetic speech

Modeling

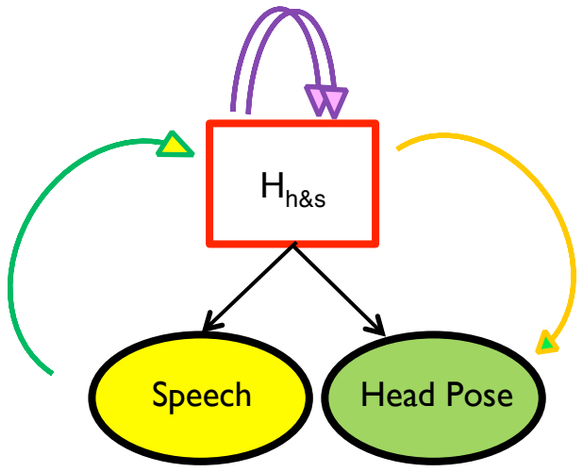
- The dynamic Bayesian network proposed by Mariooryad and Busso (2013)
- Captures the coupling between speech prosodic features and head pose
- Full observation during training
- Partial observation during testing
- Initialization by VQ





Experiments

- Three training settings:
  - C1 (Baseline):
    - Train with natural recordings
    - Mismatch
  - C2:
    - Train with the parallel corpus
    - Synthetic speech is emotionally neutral
  - C3:
    - Train with natural recording and adapt to synthetic speech
    - Mean and covariance adaptation
    - Adaptation only on speech



Adaptation

$$\mu_i = \frac{n_p \mu_{pi} + n\bar{x}_i}{n_p + n}$$

$$\sum_i = \frac{n_p\left(\sum_{pi} + (\mu_{pi} - \mu_i)^t\right) + n\overline{(x_i - \bar{x}_i)(x_i - \bar{x}_i)^t}}{n_p + n}$$

Objective Evaluation

- 5-fold cross validation
- $CCA_{s\&h}$ ↑
  - CCA between the input speech and the generated head motion sequences
- KLD ↓

- The amount of information lost by using the synthesized head movements distributions compared to the original one

	Turn-based	
	$CCA_{s\&h}$	KLD
M1	0.8615	8.4617
C1	0.8103	8.3530
C2	0.7901**	**4.7579**
C3-1	**0.8399****	8.6299
C3-2	**0.8189***	9.3203

Train & Test with original

Train with original

Train with parallel corpus

Mean adaptation

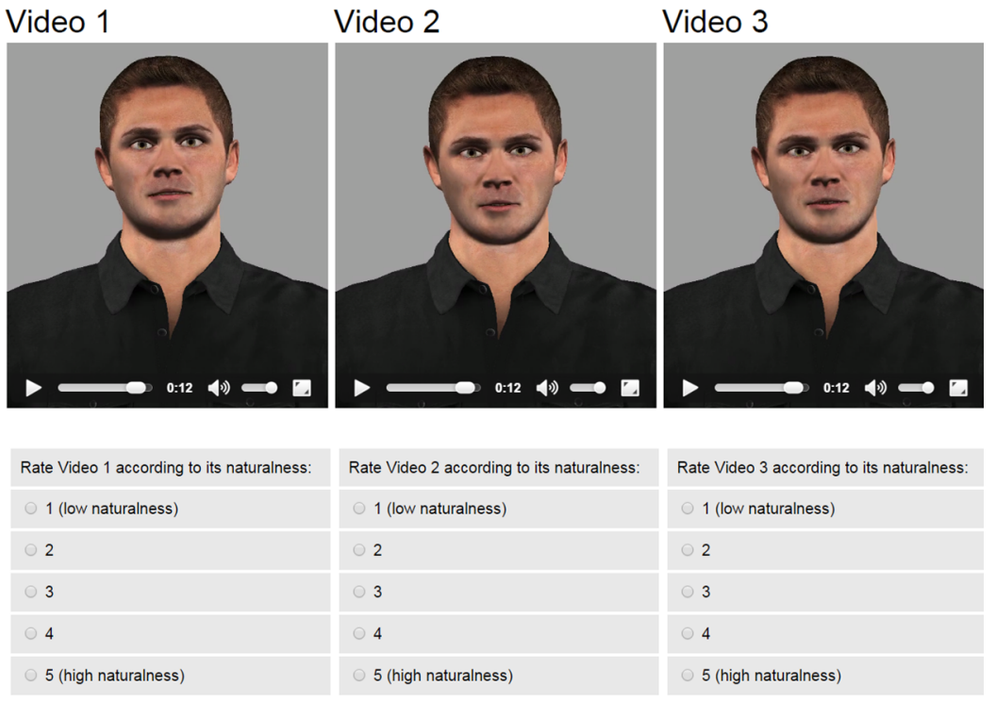
Mean & Covariance adaptation

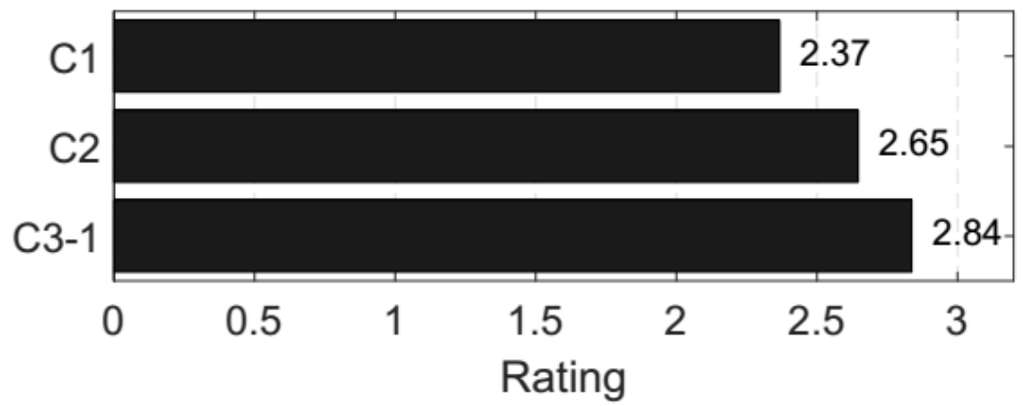
* $p < 0.05$
** $p < 0.01$

Subjective Evaluation

AMT

- Smartbody to render BVH files
- 20 videos with the three conditions (C1, C2, C3-1)
- 2 consecutive turns, to incorporate enough context
- Each evaluator is given 10 x 3 videos
- 30 evaluators in total
- Each video is annotated by 15 raters
- Kruskal-Wallis test (pairwise comparison)
  - C1 and C3-1 are different ($p < 7.4e-7$)
  - C1 and C2 are different ($p < 3.5e-3$)





Subjective Evaluation

Trained with original speech

Trained with aligned s ynthetic speech

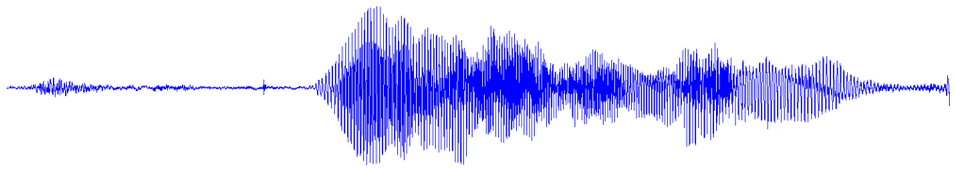
Adapted to the aligned synthetic speech

Conclusions

- This paper proposed a novel approach to scale a speech-driven framework for head motion generation to synthetic speech
- We proposed to use a corpus of synthetic speech with time-aligned signals to the natural recordings
- We used the parallel corpus to retrain or adapt the model to the synthetic speech (C2, and C3)
- This approach reduces the mismatch between train and test
- Both objective and subjective evaluations demonstrate its benefits

Future Work

- Adding emotional behaviors into our models



- Including other facial gestures (e.g., eyebrow motion) and hand gestures

- Constraining the generated behaviors on the underlying discourse function of the message to generate meaningful behaviors

Multimodal Signal Processing (MSP)

- Questions?

http://msp.utdallas.edu/