MULTIMODAL ANALYSIS OF SPEECH PROSODY AND UPPER BODY GESTURES USING HIDDEN SEMI-MARKOV MODELS

Elif Bozkurt
Multimedia, Vision and Graphics Laboratory, Koç University, Istanbul, Turkey

INTRODUCTION

- A new multimodal analysis framework for
  - Relational model of *intonation* and *gesture* phrases
  - Synthesizing gestures to emphasize speech intonation
- Requirements
  - Model correlations between modalities
  - Able to produce gesture synthesis
  - Able to model realistic gesture durations
- Our solution
  - Takes gestures as states of a Markov chain
  - Takes prosody segments as observations of this Markov process
  - Introduces a gesture duration model

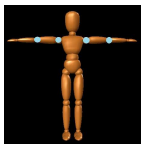
FEATURE EXTRACTION

Speech

- Prosody feature per frame
  - Normalized intensity
  - Normalized pitch
  - Pitch gain
  - Derivatives

Gesture

- Four joint angles per frame with derivatives.



UNIMODAL CLUSTERING

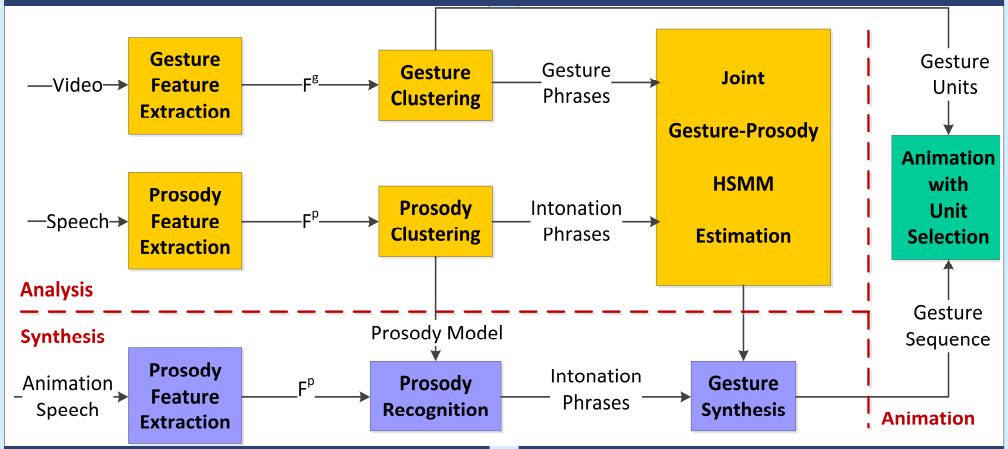
Speech

- Unsupervised clustering over prosody
- Using parallel HMMs

Gesture

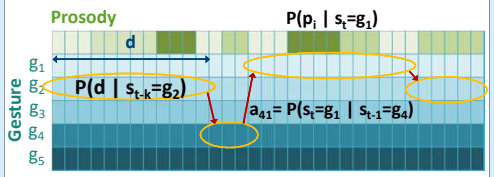
- Semi-supervised clustering over joint angles
- Using parallel HMMs
- Iterative clustering

SYSTEM OVERVIEW



MULTIMODAL ANALYSIS

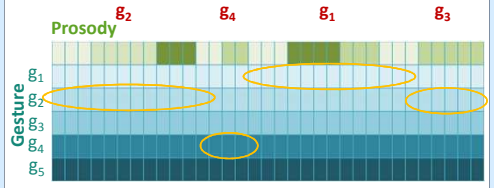
- Hidden semi-Markov model (HSMM)
  - Take gestures as *states*
  - Take prosody as the *observations*
  - Introduce a gesture duration model



Gesture Synthesis

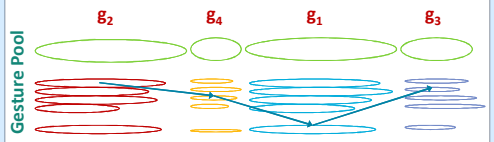
- A modified Viterbi decoder
- Forward likelihood function incorporating the state duration model

$$\psi_t(j) = \max_{\tau} \max_{i} \left\{ \psi_{t-\tau}(i) + \log\left(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(\xi_k) \right) \right\}$$



GESTURE ANIMATION

- Three main tasks
  - Motion sequence generation
    - Create a pool of gestures
    - Apply unit selection to minimize
      - i. Joint angle differences at gesture transitions
      - ii. Duration differences
  - Gesture transition smoothing
  - Gesture sequence animation



EXPERIMENTAL RESULTS

- Multimodal upper body corpus (MVGL-MUB)
  - Single subject
  - 5 recordings/totally 20 minutes.
- Leave-one-recording-out

Objective Evaluations

- Symmetric *Kullbeck-Leibler* (KL) divergence
- Minimum KL divergence for 8 prosody clusters.

States/ Branches	8	10	12	16
3	1.171	1.121	0.956	0.761
4	**0.649**	1.293	1.093	1.147
5	1.272	1.221	1.021	1.574

Subjective Evaluations

- A/B test over 9 participants
- 20 pairs of animation clips out of 72
- Clips in 40-120 seconds duration

A/B Pair	Average Preference	p-value<
Motion capture - Random	-1.130	0.0006
Motion capture – HSMM	0.019	0.9396
HSMM - Random	-0.796	0.004

CONTRIBUTIONS

- HSMM delivers a useful framework for gesture synthesis.
- Resulting animations are appealing.
- Demos are available at link http://mvgl.ku.edu.tr/icassp2013

CURRENT WORK

- Model and synthesize affective gestures
  - The USC CreativeIT database
  - Continuous modeling of speech and affect

References

- E. Bozkurt, S. Asta, S. Ozkul, Y. Yemez, and E. Erzin, "Multimodal Analysis of Speech Prosody and Upper Body Gestures using Hidden Semi-Markov Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, 2013.
- A. Metallinou, C. C. Lee, C. Busso, S. Carnicke, and S. S. Narayanan, "The USC CreativeIT Database : A Multimodal Database of Theatrical Improvisation," in *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC)*, 2010.