Audiovisual Speech Inversion by Switching Dynamical Modeling governed by a Hidden Markov Process

N. Katsamanis, G. Ananthakrishnan, G. Papandreou P. Maragos, O. Engwall

Computer Vision, Speech Communication and Signal Processing Group
http://cvsp.cs.ntua.gr
Speech Inversion

The Goal
- Identify the speech production system given observed speech

The Motives
- Understanding speech production
- Applications in speech synthesis, recognition, coding, language tutoring

The Framework
- Consider speech to be an audiovisual process

The Method
- Switching linear dynamical modeling driven by a hidden Markov process
Speech production system identification

- Describe Geometry
  - Area function, tube models
  - Articulatory models
    - Geometrical (Mermelstein 1973, Birkholz 2006)
    - Data-driven (Maeda 1979, Engwall 2003)
  - Coordinates of important articulators
    - Tongue tip, lower incisor etc.

- Determine sound sources
  - Location/Spectrum/Intensity
Data
Approaches

- From Audio only
  - Codebook (Ouni 2005), Neural networks (Richmond 2003)
  - Gaussian Mixture Model (Toda 2007), Extended Kalman Filtering (Dusan 2000), Hidden Markov Models (Hiroya 2004)

- Exploiting speaker’s facial information
  - Significant correlation between speaker’s face and vocal tract (Yehia 1998, Jiang 2002)
  - Independent component analysis of the face and relevant vector machines or neural networks to invert (Kjellstrom 2006, 2008)
  - Active appearance model for the face, canonical correlation analysis and late fusion of HMMs (Katsamanis 2007, 2008)
Contributions

- The inversion problem is one-to-many. Visual and dynamic constraints can alleviate ill-posedness. Nonlinearities can be handled efficiently in a piecewise linear manner.

- Introduction of a switching linear dynamical mechanism to model the audiovisual-to-articulatory mapping.

- Typical quantitative evaluation (RMS error) does not account for the relative importance of the errors.

- Weighted evaluation scheme based on a support vector machine classifier to determine importance of errors.
Linear Acoustic-Articulatory Mapping

- Observations $y$, vocal tract parameters $x$
  \[ p(x|y) = p(y|x)p(x)/p(y) \]

- Approximate observation model
  \[ y = Cx + \epsilon \]

- Assumptions
  \[ p(x) \sim N(x; \bar{x}, \sigma_x) \hspace{1cm} p(\epsilon) \sim N(\epsilon; 0, Q) \]

- Maximum A Posteriori
  \[ \hat{x} = (\sigma_x^{-1} + C^TQ^{-1}C)^{-1}(\sigma_x^{-1}\bar{x} + C^TQ^{-1}y) \]

- Training (Mean Square Error Minimization)
Linear Dynamic Articulatory Modeling I

- Given the observations up to moment $t$, $\mathbf{Y}_t = \{y_1, \ldots, y_t\}$

$$
p(x_t | \mathbf{Y}_t) = \frac{p(y_t | x_t)p(x_t | \mathbf{Y}_{t-1})}{p(y_t | \mathbf{Y}_{t-1})}
$$

- Analysis

$$
p(x_t | \mathbf{Y}_{t-1}) = \int p(x_t | x_{t-1})p(x_{t-1} | \mathbf{Y}_{t-1}) dx_{t-1}
$$

- Model

$$
\begin{align*}
\mathbf{x}_t &= A\mathbf{x}_{t-1} + \mathbf{w}_t & \text{Articulatory Dynamics} \\
\mathbf{y}_t &= C\mathbf{x}_t + \mathbf{v}_t & \text{Audiovisual Observation}
\end{align*}
$$

$$
\mathbf{w} \sim N(0, Q), \mathbf{v} \sim N(0, R), \mathbf{x}_0 \sim N(\mu_0, V_0)
$$
Linear Dynamic Articulatory Modeling II

■ Inference
  □ Kalman filter (MAP solution)

\[
\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(y_t - C\hat{x}_{t|t-1})
\]

■ Training/Identification
  □ State Model (Articulatory Dynamics)
    ■ Autoregressive (AR) state modeling
    ■ Maximum likelihood (MMSE)
  □ Observation Model (Audiovisual-Articulatory Mapping)
    ■ Canonical Correlation Analysis
Switching Linear Dynamic Modeling I

For a phoneme/part of a phoneme

\[ x_t = A_{1,c}x_{t-1} + A_{2,c}x_{t-2} + B_c u_c + w_t \]
\[ y_t = C_c x_t + v_t \]
\[ B_c = I - (A_{1,c} + A_{2,c}) \]

Previous work (Dusan and Deng, 2000)

- Separate model for each transition between any two phonemes
- For each model: SOM clustering to identify piecewise linear mapping
- Extended Kalman Filtering and Maximum likelihood to choose model and then Extended Kalman Smoothing

Our Assumption

- Model switching can be considered to be a Markovian process
- One model per phonemic HMM state
Switching Linear Dynamic Modeling II

Switching Process
- Audiovisual Hidden Markov Models
  - Multistream, Asynchronous

Training
- Likelihood Maximization for training (conventionally)
- Estimate responsibilities
- Train one separate Linear Dynamic System per state

Optimal State Sequence
- Phonetic Information is given
- Viterbi forced alignment
Audiovisual Speech Inversion: MOCHA

- Provided by CSTR, Univ. Edinburgh
- Two subjects (one male, one female), 460 British TIMIT Utterances each
- Articulation (2-D coords of 9 EMA coils)
- Video of the female speaker’s face
- 30 minutes of usable data
- Needed Preprocessing-labeling Video
Results

**Mean Correlation Coefficient between predicted and measured articulatory trajectories**

- **Global Kalman filter**
- **HMM**
- **Proposed Method**

For each modality:
- **Audio**
- **Visual**
- **Audiovisual**
Weighted Evaluation

LS-SVM for 2 classes, RBF kernel

- Clusters of the remaining phone classes
- Clusters of the phoneme /a/
Weighted Evaluation

Weighted Root Mean Squared (RMS) Error

\[ E_{wrms} = \frac{\sum_{k=1}^{P} \sqrt{\sum_{i \in k} (Y_i - \hat{Y}_i)^T D_k (Y_i - \hat{Y}_i)}}{N} \]

Weighting matrix for each phoneme

Classification using SVMs

Sensitivity analysis to estimate weights for each articulatory parameter
Weighted Evaluation Results

Weights for the evaluation of the inversion of plosives

- 
- 
- 
- 
- 
- 
- 
- 

lower incisor y
tongue tip y
tongue blade y
tongue dorsum y
velum y
/p/
/t/
/k/

LDS: Global Kalman Filter    HMM: Switching Linear Modeling

Proposed Method: Switching Linear Dynamic System (SLDS)

<table>
<thead>
<tr>
<th>Root Mean Square Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted</td>
</tr>
<tr>
<td>Audio</td>
</tr>
<tr>
<td>Audiovisual</td>
</tr>
</tbody>
</table>
Conclusions

- Audiovisual speech inversion framework
  - Switching linear dynamical model
  - Weighted evaluation scheme to better account for important errors

For the future

- Cope with limited data, over-fitting problems
  - Clustering
    - Tree-based (Hiroya and Honda, 2004), Data-driven
  - Adaptation
    - Adapt global regression model to local data
    - MLLR (King and Frankel, 2005) or Bayesian (Bishop, 2007) adaptation

- Invert to articulatory model parameters
Weights for the evaluation of the inversion of plosives

Thank you!