An Analysis-by-Synthesis Approach to Vocal Tract Modeling for Robust Speech Recognition

Ziad Al Bawab
(ziada@cs.cmu.edu)
Electrical and Computer Engineering
Carnegie Mellon University

Work in collaboration with:
Bhiksha Raj
Lorenzo Turicchia (MIT)
and Richard M. Stern

IBM Research
October 9, 2009
Talk Outline

I. Introduction

II. Deriving vocal tract shapes from EMA data using a physical model

III. Analysis-by-synthesis framework

IV. Dynamic articulatory model

V. Conclusion
Conventional Generative Model

SPEECH: /S/-/P/-/IY/-/CH/

Maximum Likelihood

Acoustic Features

S-P-IY-CH

Amplitude

Time

Frequency

SPEECH: /S/-/P/-/IY/-/CH/

/S/ /P/ /IY/ /CH/

S1 S2 Sn

F1 F2 ... F13

F1 F2 ... F13

F1 F2 ... F13

Nasal Cavity

Palate

Lips

Tongue

Jaw

Pharynx

Epiglottis

Larynx opening into pharynx

Larynx

Esophagus

Wikipedia
The Ultimate Generative Model

Articulatory modeling
Speech is actually generated by the vocal tract!

Physical Generative Model

Physical model of sound generation
The Missing Science

- Need a framework that can explicitly model the articulatory space (configurations and dynamics) that can help alleviate problems like coarticulation, articulatory target undershoot, asynchrony of articulators, and pronunciation variations

- Current approaches in articulatory modeling (Livescu, Deng, Erler, and more) attempt to learn and apply constraints based on inferences from surface level acoustic observations or from linguistic sources

- Need to learn from real articulatory data

- Need a mapping from articulatory space to the acoustic domain based on the physical generative process that is more natural (i.e. accurate) and can generalize better than learning the mapping statistically (i.e. from parallel articulatory and acoustic data)
MOCHA Database

MOCHA Apparatus

Raw Articulatory Measurements

Gestures
- Tongue body constrictions (extrinsic muscles)
- Tongue tip
- Tongue body shape (intrinsic muscles)
- Lip gestures

epg
- Velar port opening
- Pharyngeal constriction
- Larynx raising
- Vocal fold tension
- Glottal opening

Laryngograph

Subglottal pressure
MOCHA EMA Data

Gestures
- Tongue body constrictions (extrinsic muscles)
- Tongue tip
- Tongue body shape (intrinsic muscles)
- Lip gestures

Laryngograph
- EPG
- Velar port opening
- Pharyngeal constriction
- Larynx raising
- Vocal fold tension
- Glottal opening
- Subglottal pressure
Maeda Parameters

7 Maeda Parameters

Maeda’s Model

Area Functions (Acoustic Tubes)

<table>
<thead>
<tr>
<th>Area</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>L1</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>A36</td>
<td>L36</td>
</tr>
</tbody>
</table>
Articulatory Speech Synthesis

Area Functions (Acoustic Tubes)

<table>
<thead>
<tr>
<th>Area</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>L1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>A36</td>
<td>L36</td>
</tr>
</tbody>
</table>

Sondhi and Schroeter Model

\[
\begin{bmatrix}
P_{\text{out}} \\
U_{\text{out}}
\end{bmatrix} = \begin{bmatrix}
A & B \\
C & D
\end{bmatrix} \times \begin{bmatrix}
P_{\text{in}} \\
U_{\text{in}}
\end{bmatrix} = K \times \begin{bmatrix}
P_{\text{in}} \\
U_{\text{in}}
\end{bmatrix}
\]

VT Transfer Function
Deriving Realistic Vocal Tract Shapes from ElectroMagnetic Articulograph Data via Geometric Adaptation and Profile Fitting

- **Problem Overview:**
  - Speech synthesis solely from EMA data using:
    - Knowledge of the geometry of the vocal tract
    - Knowledge of the physics of the speech generation process

- **Approach Followed:**
  - Compute realistic vocal tract shapes from EMA data
    1. Adapting Maeda’s geometric vocal tract model to EMA data
    2. Search for best fit of the tongue and lips profile contours to EMA data
  - Synthesize speech from vocal tract shapes
    3. Articulatory synthesis using the Sondhi and Schroeter model
1. Vocal Tract Adaptation Parameters

- Origin
- Upper Wall Shift
- $\theta$
- $d$
- Lips Separation
Adaptation Result [1]

2. Search Results

(a) ‘II’
EMA points in purple for phone ‘II’ as in “Seesaw = /S-II-S-OO/”

(b) ‘@@’
EMA points in purple for phone ‘@@’ as in “Working = /W-@@-K-I-NG/”
3. Synthesis Results

(a) ‘II’
Acoustic tubes model for phone
‘II’ as in “Seesaw = /S-II-S-OO/”

(b) ‘@@’
Acoustic tubes model for phone
‘@@’ as in “Working = /W-@@-K-I-NG/”
Creating a Realistic Codebook and Adapted Articulatory Transfer Functions

EMA → Search Adapted Maeda's Model → Adapted Vocal Tract Shapes → Articulatory Synthesis Model → Adapted Transfer Functions

Velum Area

Codeword: p1 p2 p3 p4 p5 p6 p7 VA
Projecting the 44 Phones Codewords’ Means using Multi-Dimensional Scaling (MDS)
Deriving Analysis-by-Synthesis Features[2]

Compare signals generated from a codebook of valid vocal tract configurations to the incoming signal to produce a “distortion” feature vector.

Articulatory Space

Articulatory Configurations

codeword 1
  P1
  P7

codeword N
  P1
  P7

Energy, Pitch

Synthesis

MFCC

Mel-Cepstral Distortion

Speech

Distortion Feature Vector

Mixture Probability Density Function

• For a given frame, the output probability of each state in the HMM is a mixture density over a set of M codewords:

\[
P(x|S_1) = \sum_{j=1}^{M} P(x, cd_j|S_1) \\
= \sum_{j=1}^{M} \frac{P(cd_j|S_1)}{\sum_{k=1}^{M} P(cd_k|S_1)} P(x|cd_j, S_1) \\
= \sum_{j=1}^{M} w_{1j} P(x|cd_j, S_1)
\]
HMM Framework

\[ bd_i(t) = P(x(t) | S(t) = i) \]

\[ = \sum_{k=1}^{M} w_{ik} \lambda_{ik} \exp^{-\lambda_{ik} d_k^2(t)} \]
$w_j = \frac{\text{count\_frames}(\text{phone} = C, \text{truecode} = \text{cd}_j)}{\text{total\_frames}(\text{phone} = C)}$  

\[ j = 1 : M \]

EMA measurements

Time

TT
TB
TD

\(\text{cd}_1\) \hspace{1cm} \text{cd}_2 \hspace{1cm} \text{cd}_1 \hspace{1cm} \text{cd}_3 \hspace{1cm} \text{cd}_2\)
Update Equations

• For each phone, we estimate \( \{w_1, w_2, \ldots, w_M\} \) and \( \{\lambda_1, \lambda_2, \ldots, \lambda_M\} \) for each state as:

\[
\begin{align*}
  w_j^t &= \frac{1}{U} \sum_{u=1}^{U} P(c_d_j | x_u, \theta_j^{t-1}) \\
  &= \frac{1}{U} \sum_{u=1}^{U} \frac{w_j^{t-1} P(x_u | c_d_j, \theta_j^{t-1})}{\sum_{k=1}^{M} w_k^{t-1} P(x_u | c_d_k, \theta_k^{t-1})}
\end{align*}
\]

\[
\begin{align*}
  \lambda_j^t &= \frac{\sum_{u=1}^{U} P(c_d_j | x_u, \theta_j^{t-1})}{\sum_{u=1}^{U} d_{u,j}^2 P(c_d_j | x_u, \theta_j^{t-1})} \\
  &= \frac{\sum_{u=1}^{U} w_j^{t-1} P(x_u | \lambda_j^{t-1})}{\sum_{k=1}^{M} \sum_{u=1}^{U} w_k^{t-1} P(x_u | \lambda_k^{t-1})} \\
  &= \frac{\sum_{u=1}^{U} d_{u,j}^2 w_j^{t-1} P(x_u | \lambda_j^{t-1})}{\sum_{k=1}^{M} \sum_{u=1}^{U} d_{u,j}^2 w_k^{t-1} P(x_u | \lambda_k^{t-1})}
\end{align*}
\]
Weights for Phone ‘OU’ Projected on Codewords-MDS Space

Priors from EMA

Weights Flat Init

Weights Init from EMA

Weights Init from EMA + Adaptation
Experimental Setup

• Segmented phone recognition on the MOCHA Database (9 speakers, 460 TIMIT British English utterances per speaker, 44 phones)

• Articulatory codebook composed of 1024 different Maeda configurations derived from MOCHA EMA data

• LDA dimensionality reduction of the distortion vector to 20 features per frame, phones being the classes of transformation
Experimental Setup Cont’d

• Distortion measure used is the Mel-Cepstral distortion:

\[
MCD(\vec{C}_{\text{incoming}}, \vec{C}_{\text{synth}}) = \frac{10}{\ln 10} \sqrt{\sum_{k=1}^{12} (C_{\text{incoming}}(k) - C_{\text{synth}}(k))^2}
\]

• Classify each phone c according to:

\[
\hat{c} = \arg \max_c P(c)P(MFCC | c)^\alpha P(DF | c)^{1-\alpha}
\]
## Summary of Phone Error Rates Results [3]

<table>
<thead>
<tr>
<th>Features (dimension)</th>
<th>Topology</th>
<th>fsew0 14,352</th>
<th>msak0 14,302</th>
<th>Both 28,654</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC + CMN (13)</td>
<td>3S-128M-HMM Gaussian/VQ</td>
<td>61.6%</td>
<td>55.9%</td>
<td>58.8%</td>
<td></td>
</tr>
<tr>
<td>DistFeat (1024)</td>
<td>3S-1024M-HMM Exponential/Flat Sparsity = 21%</td>
<td>57.6%</td>
<td>53.7%</td>
<td>55.7%</td>
<td>5.3%</td>
</tr>
<tr>
<td>DistFeat (1024)</td>
<td>3S-1024M-HMM Exponential/EMA Sparsity = 51%</td>
<td>58.3%</td>
<td>53.9%</td>
<td>56.1%</td>
<td>4.6%</td>
</tr>
<tr>
<td>Adapted DistFeat (1024)</td>
<td>3S-1024M-HMM Exponential/EMA Sparsity = 51%</td>
<td>58.4%</td>
<td>53.1%</td>
<td>55.7%</td>
<td>5.3%</td>
</tr>
<tr>
<td>DistFeat + LDA + CMN (20)</td>
<td>3S-128M-HMM Gaussian/VQ Sparsity = 0%</td>
<td>54.9%</td>
<td>49.8%</td>
<td>52.4%</td>
<td>10.9%</td>
</tr>
</tbody>
</table>

## Summary of our Contribution

<table>
<thead>
<tr>
<th></th>
<th>Conventional HMM</th>
<th>Production Based HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>States</strong></td>
<td>Abstract, no physical meaning</td>
<td>Real articulatory configurations</td>
</tr>
<tr>
<td><strong>Output Observation</strong></td>
<td>Gaussian probability using acoustic features</td>
<td>Exponential probability based on the analysis-by-synthesis distortion features</td>
</tr>
<tr>
<td><strong>Probability</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Adaptation</strong></td>
<td>VTLN, MLLR, MAP</td>
<td>Vocal tract geometric model adaptation</td>
</tr>
<tr>
<td><strong>Transition Probability</strong></td>
<td>Based on acoustic observation</td>
<td>Can be leaned from articulatory dynamics</td>
</tr>
</tbody>
</table>
Conclusion

• A model that mimics the actual physics of the vocal tract results in better classification performance

• Developed a hybrid physical and statistical dynamic articulatory framework that incorporates analysis-by-synthesis for improved phone classification

• Recent databases open new horizons to better understand the articulatory phenomena

• Current advancements in computations and machine learning algorithms facilitate the integration of physical models in large scale systems
• THANK YOU 😊