AERFAI Summer School

Speech Production Models in ASR

Richard Rose
June, 2008

McGill University
Dept. of Electrical and Computer Engineering
OUTLINE

1. Speech Production Models
   – Motivating Articulatory Based Models for ASR
   – Review of Speech Production and Distinctive Features
   – Sounds to Words – Problems with Pronunciation Dictionaries
   – The Role of Speech Production Models in Speech Perception

2. Exploiting Speech Production Models in ASR
   – Statistical methods for phonological distinctive feature detection
   – Incorporating distinctive feature knowledge in ASR model structure
   – Development of models of articulatory dynamics
   – Integrating distinctive features in traditional ASR systems

3. Resources for Research
   – Articulatory measurements and clinical tools
   – Speech corpora
   – Projects dedicated to speech production models in ASR
1. Speech Production Models

- Motivating Articulatory Based Models for ASR

- Review of Speech Production and Distinctive Features

- Sounds to Words – Problems with Pönemic Pronunciation Dictionaries

- The Role of Speech Production Models in Speech Perception
Motivating Articulatory-Based Models for ASR

• A case for Articulatory Representations
  – Speech as an organization of articulatory movements
  – Critical articulators – Invariance in the articulatory space
  – Evidence for usefulness of articulatory knowledge
The Organization of Articulatory Movements

- Speech production can be described by the motion of loosely synchronized articulatory gestures.

- Motivates the use of multiple streams of semi-independent phonological features in ASR.

- Suggests that segmental, phonemic models are problematic.

Acoustic waveform and measured articulatory trajectories for utterance of “It’s a /bamib/ sid” (Krakow, 1987)
Reduced Variability Through Critical Articulators

- ASR models with structure defined in an articulatory domain may exploit invariance properties associated with critical articulators

- Critical Articulator: “The articulator most crucially involved in a consonant production”

- Less susceptible to coarticulatory influences

- Less overall variability

Peek-to-Peak X-ray microbeam Trajectories

Papcun et al, 1992
Evidence for Usefulness of Articulatory Information

- ASR Performance Improved using “direct measurements”
  - Electromagnetic Articulography (EMA) [Zlokarnik, 1993][Wrench, 2002]

<table>
<thead>
<tr>
<th>Acoustic ASR</th>
<th>65%</th>
<th>89.4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic + Art</td>
<td>78%</td>
<td>94.4%</td>
</tr>
</tbody>
</table>

Sequence of binary lip/tongue images for word “one” (Petajan)

Placement of EMA coils (Zlokarnik)
“Partial” Direct Measurements - Visual Information

• Partial direct articulatory measurements fused with acoustic information in audio-visual ASR [Potamianos et al, 2004]
Motivating Articulatory-Based Models for ASR

• Challenges for Incorporating Articulatory Models
  – One-to-many acoustic to vocal tract area mapping
  – Non-linear relationship between production, acoustics, and perception
  – Coding of perceptually salient articulatory information
Acoustic to Vocal Tract Area Mapping

- Mapping from transfer function to area function is not unique
- Inversion techniques affected by source excitation

![Diagram of Vocal Tract Area Functions and Frequencies, Bandwidths]

<table>
<thead>
<tr>
<th></th>
<th>653</th>
<th>1188</th>
<th>2177</th>
<th>3321</th>
<th>HZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>132</td>
<td>119</td>
<td>188</td>
<td></td>
<td>HZ</td>
</tr>
</tbody>
</table>

Frequencies, Bandwidths

<table>
<thead>
<tr>
<th></th>
<th>653</th>
<th>1188</th>
<th>2177</th>
<th>2783</th>
<th>143</th>
<th>HZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>139</td>
<td>169</td>
<td>143</td>
<td></td>
<td></td>
<td>HZ</td>
</tr>
</tbody>
</table>
Acoustic Coding of Articulatory Information

- Perceptually salient information necessary for making phonemic distinctions can be contained in fast-varying, short duration acoustic intervals [Furui, 1986]
- Difficult to exploit this information to predict motion of articulators
- Evidence: Japanese CV syllable identification tests [Furui, 1986]

[From Furui, 1986]
1. Speech Production Models

- Motivating Articulatory Based Models for ASR

- Review of Speech Production and Distinctive Features

- Sounds to Words – Problems with Pronunciation Dictionaries

- The Role of Speech Production Models in Speech Perception
A Brief Review of Distinctive Features

- We need a way to describe the sounds of speech in any language in terms of the underlying speech production system.

- **Distinctive Features** – Serve to distinguish one phoneme from another by describing:
  
  1. **The Manner** in which the sound is produced
     - Voiced, Unvoiced, Vocalic, Consonantal, Nasal
  
  2. **The Place** where the sound is articulated
     - Labial, Dental, Alveolar, Palatal, Velar
Speech Production – Distinctive Features

HUMAN VOCAL SYSTEM

Nasal Cavity: Produces sound when velum is open and air flows through the nose

Articulators: Tongue, Lips, Teeth, Velum, and Hard Palate

Vocal Folds / Glottis: Vibrates in response to air pressure originating in the lungs

[from Rabiner and Juang, 1993]
Speech Production – Distinctive Features

- Manner of Production

  - Voiced: Glottis closed with glottal folds vibrating
  - Unvoiced: Glottis open
  - Sonorant: No major constriction in the vocal tract and vocal cords set for voicing
  - Consonantal: Major constriction in vocal tract
  - Nasal: Air travels through the nasal cavity

[from Rabiner and Juang, 1993]
Speech Production – Distinctive Features

HUMAN VOCAL SYSTEM

• Place of Articulation

  • Bilabial - Lips - /P/, /B/, /M/

  • Dental - Tongue Tip and Front Teeth - /TH/, /DH/

  • Alveolar - Alveolar Ridge and Tip of Tongue - /T/, /D/, /N/, /S/, /Z/, /L/

  • Palatal - Hard Palate and Tip of Tongue - /Y/, /ZH/

  • Velar - Soft Palate (Velum) and Back of Tongue - /K/, /G/, /NG/

[from Rabiner and Juang, 1993]
Classes of Sounds: Vowels

• Distinctive Features that are common to all vowels:
  +Voiced, +Sonorant, -Consonantal

• Vowels are distinguished by Distinctive Features:
  • Tongue Position: Front, Mid, Back
  • Jaw Position: High, Mid, Low
  • Lip Rounding: Rounded, Not-Rounded
  • Tense / Lax: Widening of the cross-sectional area of the pharynx by moving the tongue root forward
## Vowels of English

English vowels include monothongs, diphthongs, and reduced vowels

### TONGUE BODY

<table>
<thead>
<tr>
<th>JAW POSITION</th>
<th>Front</th>
<th>Mid</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>/IY/ peat</td>
<td>/ER/ pert</td>
<td>/UW/ boot</td>
</tr>
<tr>
<td></td>
<td>/IH/ pit</td>
<td></td>
<td>/UH/ foot</td>
</tr>
<tr>
<td>Mid</td>
<td>/EY/ eight</td>
<td>/AH/ putt</td>
<td>/OH/ open</td>
</tr>
<tr>
<td></td>
<td>/EH/ pet</td>
<td></td>
<td>/AO/ all</td>
</tr>
<tr>
<td>Low</td>
<td>/AE/ pat</td>
<td></td>
<td>/AA/ father</td>
</tr>
</tbody>
</table>

**tense / lax pairs**

**REDUCED VOWELS:** /AX/ about /IX/ roses /AXR/ butter

**DIPTHONGS:** /AY/ bite /OY/ Boyd /AW/ bout
Classes of Sounds: Consonants

• Distinctive Features that are common to all consonants:
  - Sonorant, +Consonental

• Consonants are distinguished by distinctive features:
  • Place of Articulation
    • Labial, Dental, Aveolar, Palatal, Velar
  • Manner of Articulation
    • Stop: Complete Stoppage of airflow in the Vocal Tract followed by a release
    • Fricative: Noise from constriction in the vocal tract
    • Nasal: Velum open and air flows through nasal cavity
Classes of Sounds: Fricatives

/F/
find
Labial

/TH/
the
Dental

/S/
say
Alveolar

/SH/
show
Palatal-Alveolar
Classes of Sounds: Nasals and Affricatives

- **Nasals:**
  - Distinctive Feature Common to Nasals is +nasal (velum open)
  - Distinguished by places of articulation
    - /M/ mom – labial
    - /N/ none – alveolar
    - /NG/ sing - velular

- **Affricatives:**
  - Alveolar-stop palatal-fricative pair
  - Distinguished by voicing
    - /JH/ judge – voiced
    - /CH/ church – unvoiced

- **Aspirant:**
  - One aspirant in English produced by turbulent excitation at the glottis
    - /H/ hat
Classes of Sounds: Semi-Vowels

- Transition Sounds:

  - Liquids: Some obstruction of the airstream of the mouth but not enough to cause frication

    /L/ - lack       /R/ - red

  - Glides: Tongue moves rapidly in a gliding fashion either toward or away from neighboring vowel

    /W/ - way       /Y/ - you
Example: Distinctive Features used to Define Phonological Rules for Morphologically Related Words

An example: The plural form of English nouns

- Orthographically: Plural is formed by adding “s” or “es”
- Phonemically: Plurals result in adding one of three endings to the word: /S/, /Z/, or /IH/ /Z/
- The actual ending depends on the last phoneme of the word.

Which plural ending would be associated with the following 3 groups of words?

What is the minimum feature set for the phonemes that proceed these plural endings?

1. breeze, fleece, fish, judge, witch
   /IH//Z/: +consonental, +strident, -stop, +alveolar

2. mop, lot, puck, leaf, moth
   /S/: +consonental -vocalic -voiced

3. tree, tray, bow, bag, mom, bun, bang, ball, bar
   /Z/: +voiced
Phonology: From Phonemes to Spoken Language

- **Phonology**: Mapping from baseform phonemes to acoustic realizations (surface form phonemes)
- **Allophones**: Predictable phonetic variants of a phoneme
- **Phonological Rules**: Applied to phoneme strings to produce actual pronunciation of words in sentences
  - **Assimilation**: Spreading of phonetic features across phonemes
  - **Flapping**: Change alveolar stop to a “flap” when spoken between vowels
  - **Nasalization**: Impart nasal feature to vowels preceding nasals
  - **Vowel Reduction**: Change vowel to /AX/ when unstressed

<table>
<thead>
<tr>
<th>Representations</th>
<th>Flapping Rule (CITY)</th>
<th>Vowel Reduction Rule (PHONOLOGY)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonemic</td>
<td>/C/ /IH/ /T/ /IY/</td>
<td>/F/ /OH/ /N/ /AA/ /L/ /AX/ /J/ /IY/</td>
</tr>
<tr>
<td>Phonetic</td>
<td>/C/ /IH/ /D/ /IY/</td>
<td>/F/ /AX/ /N/ /AA/ /L/ /AX/ /J/ /IY/</td>
</tr>
</tbody>
</table>
1. Speech Production Models

- Motivating Articulatory Based Models for ASR
- Review of Speech Production and Distinctive Features
- Sounds to Words – Problems with Phonemic Pronunciation Dictionaries
- The Role of Speech Production Models in Speech Perception
Sounds to Words – Problems with Dictionaries

Mismatch: Canonical baseforms vs. Surface Form Variant

- Surface-form phone models can be trained using surface acoustic trans.: 

\[
\begin{align*}
\text{Word} \quad W & \quad \rightarrow \quad \{p_1, \ldots, p_k, \ldots, p_N\} \\
\text{Canonical Phonetic Baseform} & \quad \rightarrow \quad \text{Pron. Variant 1} \\
\text{Acoustic Space} & \quad \rightarrow \quad \lambda_k^1 \\
\text{Surface Transcriptions} & \quad \rightarrow \quad \{\lambda_k^1, \lambda_k^2\} \\
\text{Phone Models} & \quad \rightarrow \quad \lambda_k^2
\end{align*}
\]

- The challenge is to predict pronunciation variants during recognition:

\[
p_k \rightarrow \{\lambda_k^1, \lambda_k^2\}
\]
## Problems with Dictionaries

### Base-form vs. surface-form pronunciations:

<table>
<thead>
<tr>
<th>Word</th>
<th>Purpose</th>
<th>and</th>
<th>Respect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base-Form</td>
<td>perp - axs ade nd r ih s p - eh k t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surface-Form</td>
<td>prerp pcl prix s eh n - r ix s pcl pr eh kcl tr</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Canonical Pronunciation Dictionary

**Coverage vs. Ambiguity**

- Adding pronunciation variants to increase coverage can introduce ambiguity among dictionary entries

<table>
<thead>
<tr>
<th>Word</th>
<th>Canonical Baseform</th>
</tr>
</thead>
<tbody>
<tr>
<td>an</td>
<td>/eh/ /n/</td>
</tr>
<tr>
<td>and</td>
<td>/ae/ /n/ /d/</td>
</tr>
<tr>
<td>had</td>
<td>/h/ /ae/ /d/</td>
</tr>
<tr>
<td>head</td>
<td>/h/ /eh/ /d/</td>
</tr>
<tr>
<td>purpose</td>
<td>/p/ /er/ /p/ /ax/ /s/</td>
</tr>
<tr>
<td>respect</td>
<td>/r/ /ih/ /s/ /p/ /eh/ /k/ /t/</td>
</tr>
</tbody>
</table>
Impact of Canonical Phonemic Baseforms

• **Speaking Style: Increased speaking rate** [Bernstein et al, 1996]
  – Number of words per second increases with speaking rate
  – Number of phones per second stays roughly the same
  – Phones are deleted, not just reduced

• **Speaking Style: Spontaneous Speech** [Fosler et al, 1996]
  – Switchboard Corpus: ~67% of labeled phones agree with canonical pronunciations

• **Inherent Ambiguity of the Phoneme** [Greenberg, 2000]
  – Inter-labeler agreement for labeling phonemes in spontaneous speech is only 75 to 80 percent

**Potential:** Huge WAC improvement possible

ASR with “Correct Pronunciations” can increase WAC by 40%
Impact of Canonical Phonemic Baseforms

• Better modeling of surface-form phones does not increase WAC

• Demonstration: TIMIT Corpus
  – Train context dependent HMM phone models from
    • Surface-form (S-F) acoustic transcriptions – manually labeled
    • Base-form (B-F) transcriptions – From canonical pronunciations

<table>
<thead>
<tr>
<th>Word Trans.</th>
<th>purpose</th>
<th>and</th>
<th>respect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base-Form Trans.</td>
<td>p  er</td>
<td>p  ax</td>
<td>s  ae</td>
</tr>
<tr>
<td>Surface-Form Trans.</td>
<td>p  er</td>
<td>p  ix</td>
<td>s  ix</td>
</tr>
</tbody>
</table>

– Compare phone accuracy (PAC) and word accuracy (WAC) using S-F and B-F HMM models [Rose et al, 2008]
Impact of Canonical Phonemic Baseforms

- Better modeling of surface-form phones does not increase WAC

- Demonstration: TIMIT Corpus
  - Train context dependent HMM phone models from
    - Surface-form (S-F) acoustic transcriptions – manually labeled
    - Base-form (B-F) transcriptions – From canonical pronunciations
  - Phone accuracy (PAC) and word accuracy (WAC) [Rose et al, 2008]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface-form</td>
<td>69.1%</td>
<td></td>
<td>92.0%</td>
</tr>
<tr>
<td>Base-form</td>
<td></td>
<td>63.3%</td>
<td>96.1%</td>
</tr>
</tbody>
</table>

- HMMs trained from S-F trans. provide best model of acoustic variants

  … But this does not result in better ASR word accuracy
1. Speech Production Models

- Motivating Articulatory Based Models for ASR
- Review of Speech Production and Distinctive Features
- Sounds to Words – Problems with Pronunciation Dictionaries
- The Role of Speech Production Models in Speech Perception
Connection Between Distinctive Features and Speech Perception

- **Quantal Theory of Speech Perception:** Every distinctive feature in every language represents a nonlinear discontinuity in the relationship between articulatory position and acoustic output [Stevens, 1989]

  ![Diagram](image)

- **Example:** Opening velum by $T_2 - T_1 = 2$ millimeters while uttering the phoneme /d/ causes an increase in acoustic output energy of 20 – 30 dB
  - /d/ becomes /n/ and [-sonorant] becomes [+sonorant]
  - Similar non-linear discontinuities exist in the relationship between acoustics and perceptual space
A Model of Human Speech Perception - Distinctive Features and Acoustic Landmarks

- Model speech perception process using a discrete lexical representation [Stevens, 2002]:
  - Words are a sequence of discrete segments
  - Segments are a discrete set of distinctive features

- Landmarks: Provide evidence for broad classes of consonant or vowel segments

- Articulatory Features: Associated with articulation event and acoustic pattern occurring near landmarks
Landmark / Feature Based Model of Human Perception

Model of Lexical Access in Human Speech Perception [Stevens, 2002]

- Vowel Landmarks - Peaks in first formant
- Consonant Landmarks - Acoustic discontinuities
- Articulator Bound Features – Extracted from Acoustic Cues within tens of milliseconds of landmarks
- Words in Lexicon – Formed from segments made up of “bundles” of features

[From Stevens, 2002]
Landmark / Feature Based Model of Human Perception

Model of Lexical Access in Human Speech Perception [Stevens,2002]

Analysis-by-Synthesis: Incorporating higher level linguistic knowledge for re-evaluating hypothesized word sequences [Stevens,2000]
2. Exploiting Speech Production Models in ASR

- Statistical methods for phonological distinctive feature (PDF) detection

- Incorporating distinctive feature knowledge in ASR model structure

- Articulatory models of vocal tract dynamics

- Integrating distinctive features in traditional ASR systems
Statistical methods for phonological distinctive feature (PDF) detection

• The definition of PDFs for ASR
• Obtaining acoustic parameters from surface acoustic measures
• Issues for incorporating PDFs and training PDF Detectors
• Statistical methods for PDF detection
Phonological Distinctive Features (PDFs) for ASR

• Few ASR systems exploit direct Articulatory Measurements
  – Other examples - low power radar sensors (GEMS) [Fisher,2002]

• Many ASR systems exploit phonological distinctive features

• PDFs used as a “hidden process”
  – Exploit advantages of articulatory based representation
  – Overlapping, as opposed to segmental, models of speech
  – Invariance properties associated with critical articulators
Phonological Distinctive Features (PDFs) for ASR

• Example of multi-valued definition of PDFs [King et al, 2000]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manner of Articulation</td>
<td>Vowel, Fricative, Approximant, Nasal</td>
</tr>
<tr>
<td>Place of Articulation</td>
<td>Low, Mid, High, Palatal, Labial, Coronal-Dental, Labial-dental, Labial, Coronal, Velar, Glottal …</td>
</tr>
<tr>
<td>Phonation</td>
<td>Voiced, Unvoiced</td>
</tr>
<tr>
<td>Centrality</td>
<td>Central, Full, Undefined</td>
</tr>
<tr>
<td>Continuant</td>
<td>Continuant, Non-continuant</td>
</tr>
<tr>
<td>Front-back</td>
<td>Back, Front</td>
</tr>
<tr>
<td>Roundness</td>
<td>Round, Not-Rounded</td>
</tr>
<tr>
<td>Tenseness</td>
<td>Lax, Tense</td>
</tr>
</tbody>
</table>

• Many other definitions of Features
  – Binary PDFs [Chomsky and Halle, 1967]
  – Government Phonology [Haegeman, 1994][Ahern, 1999]
  – Articulatory Features [Deng and Sun, 1999] [Bridle et al, 1998]
Phonological Distinctive Features (PDF) for ASR

- Obtaining Acoustics Correlates of PDFs from Surface Acoustic Waveforms
  - Acoustic Correlates: Relationship between S-A parameters and PDFs

![Diagram](image.png)
Obtaining PDF’s from Surface Acoustic Measures

- Define acoustic correlates for a feature
- Determine acoustic parameters that characterize acoustic correlates
  - Example: acoustic parameters for stop consonants [Epsy-Wilson]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Acoustic Correlates</th>
<th>Acoustic Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop consonant (non-continuant)</td>
<td>Closure followed by abrupt spectral change</td>
<td>Closure:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Energy: 0.2-3KHz</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Energy: 3-6KHz</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACorr: R(1)/R(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Burst: Spectral Flatness</td>
</tr>
</tbody>
</table>

- Acoustic parameters and feature detectors
  - Feature space transformations (LDA) and feature selection algorithms allow acoustic parameters to be identified from candidate params.
Phonological Distinctive Features (PDF) for ASR

- Detecting PDFs from Acoustic Parameters
  - Non-linear relationship between acoustic and articulatory distances

![Diagram of speech production models in ASR](image-url)
Issues for Training Statistical PDF Detectors

• **Supervised Training** – Defining “True” Feature Labels in Training
  – Mapping from phone to feature transcriptions [King et al., 2000]
    - Actual feature values may differ from canonical values
  – Using direct physical measurements [Wrench et al., 2000]
    - Difficult to convert physical measurements to feature values
  – Manual labeling of distinctive features [Livescu et al., 2007]
    - Defining labeling methodology, Time consuming (~1000 times RT)
  – Embedded Training – Allow feature boundaries to vary [Frankel et al., 2007]
    - Provides re-alignment of features, but no measure of quality
Detecting PDFs From Surface Acoustic Parameters

• Relationship between articulatory distances and acoustic distances can be highly nonlinear [Niyogi et al, Stevens et al]

• Only small regions of acoustic space correspond to regions of high articulatory discriminability

• Fits nicely as a problem for support vector machines (SVM)

Nonlinear PDF Detectors:
- SVM [Niyogi et al]
- TDNN [King and Taylor]
- MLP [Kirchhof]
Detecting PDFs From Surface Acoustics – Dynamic Bayesian Networks

- **Modeling Asynchrony Among Distinctive Features**
  - Dynamic Bayes networks (DBN) [Frankel et al, 2007][Livescu et al, 2004]

![Diagram of Dynamic Bayesian Networks]

Continuous, observable acoustic variables

\[ P(Y_t | X_t^k) \]

Discrete, hidden Distinctive feature variables

Dependencies between features encoded by conditional probabilities

\[ P(X_t^k | X_t^1, ..., X_t^N, X_{t-1}^k) \]

(Model of PDF Dynamics)

-Manner
-Voicing
-Place
-Front/Back
-Static
-Rounding

[From Frankel et al, 2007]
Detecting PDFs Using Dynamic Bayesian Networks

• **Modeling Acoustic Observations** $P(Y_t | X_t^k)$: Gaussian mixtures or artificial neural networks

• **Modeling PDF State Process** $P(X_t^k | X_t^1, .., X_t^N, X_{t-1}^1, .., X_{t-1}^N)$: Hierarchical conditional probability tables – Allows for asynchrony among feature values

• **Embedded Training:**
  – Initial training performed using phone alignments converted to feature values
  – Generate new PDF alignments and retrain with re-aligned transcriptions

• **Effects on Phone Recognition Accuracy:**
  – Frankel et al found that embedded training had very little effect on phone accuracy [Frankel, 2007]
  – Observed feature asynchrony was representative of speech production
2. Exploiting Speech Production Models in ASR

• Statistical methods for phonological distinctive feature (PDF) detection

• Incorporating distinctive feature knowledge in ASR model structure

• Development of models of articulatory dynamics

• Integrating distinctive features in traditional ASR systems
ASR Model Structure Based on PDFs

- **A Case for Model Structure Based on PDFs**
  - HMM State Space: Model topology defined by feature spreading
  - Pronunciation: Feature based description of pronunciation variation
  - A Complete Model: Implementation of landmark based / distinctive feature approach to ASR
Modeling Structure Based on PDF’s

- **PDF Based HMM state space** [Deng and Sun, 1999]
  - Phones in context defined in terms of articulatory features
  - Context specific nodes formed by spreading features
  - PDF based nodes permit defining context in articulatory space

Phone in Context Models – State Trans. Graphs

HMM States defined as Multi-valued Articulatory Features

- Lips
- Tongue Body
- Tongue Dorsum
- Velum
- Larynx

Left influence of TB value 1

Right influence of TD value 9
Modeling Structure Based on PDF’s

- **PDF based models of pronunciation variation** [Livescu et al, 2004]
  - PDFs model asynchrony of articulators and articulatory dynamics
  - Model structure based on dynamic Bayesian networks (DBNs)

- **Canonical Dictionary Expanded as PDFs** [Livescu et al, 2004]

<table>
<thead>
<tr>
<th>Word</th>
<th>and</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phones</td>
<td>ae</td>
</tr>
<tr>
<td>Index</td>
<td>0</td>
</tr>
<tr>
<td>Phonation</td>
<td>Voiced</td>
</tr>
<tr>
<td>Manner</td>
<td>Vowel</td>
</tr>
<tr>
<td>Place</td>
<td>Low</td>
</tr>
<tr>
<td>Continuant</td>
<td>Continuant</td>
</tr>
</tbody>
</table>
### Canonical Articulatory Baseforms

- **Canonical Dictionary Expanded as PDFs** [Livescu et al, 2004]

<table>
<thead>
<tr>
<th>PDF</th>
<th>Baseform</th>
<th>Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word and</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phones</td>
<td>ae</td>
<td>n</td>
</tr>
<tr>
<td>Index</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Phonation</td>
<td>Voiced</td>
<td>Voiced</td>
</tr>
<tr>
<td>Manner</td>
<td>Vowel</td>
<td>Nasal</td>
</tr>
<tr>
<td>Place</td>
<td>Low</td>
<td>Coronal</td>
</tr>
<tr>
<td>Continuant</td>
<td>Continuant</td>
<td>Non-Continuant</td>
</tr>
</tbody>
</table>

- **Probabilistic Models of Feature Asynchrony and Feature Substitution**

**Articulatory Asynchrony**

<table>
<thead>
<tr>
<th>Articulatory Asynchrony</th>
<th>Manner Index</th>
<th>Place Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Articulatory Dynamics**

<table>
<thead>
<tr>
<th>Articulatory Dynamics (Feature Substitution)</th>
<th>Underlying $U_t^i$</th>
<th>Observed $X_t^i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vow</td>
<td>Vow</td>
<td>Vow</td>
</tr>
<tr>
<td>Vow</td>
<td>Vow</td>
<td>Nas</td>
</tr>
</tbody>
</table>

**Asynchrony Model:**

$$P(|\text{Index}(X_t^i) - \text{Index}(X_t^j)|)$$

**Substitution Model:**

$$P(X_t^i = x | U_t^i = y)$$

Feature Frames (t) → ...
Landmark / Feature Based Model of Human Perception

Model of Lexical Access in Human Speech Perception [Stevens, 2002]

- Vowel Landmarks - Peaks in first formant
- Consonant Landmarks - Acoustic discontinuities
- Articulator Bound Features – Extracted from Acoustic Cues within tens of milliseconds of landmarks
- Words in Lexicon – Formed from segments made up of “bundles” of features

[From Stevens, 2002]
Landmark / Distinctive Feature Based Approach to ASR

Landmark-Based Speech Recognition
[Hasegawa-Johnson et al, 2005]

Speech

Extract Acoustic Correlates of Features

Acoustic Correlates: \( Y_t \)

SVM Based Detector 1

SVM Based Detector 72

Posters: \( P(d_j(t) | Y_t, L_t) \)

Dynamic Programming Based Landmark Detection

Lexicon

Baseline ASR Lattices

Lattice Rescoring

Hypothesized Word Sequences

• Acoustic Parameters:

• Distinctive Feature Detectors:
  – Support Vector Machines (SVMs)
  – Produce posterior probabilities of distinctive feature values \( d_j(t) \) for landmark type \( L_t \) at time \( t \).

• Landmark Detection
  – Maximizes posterior probability of distinctive feature “bundles” w.r.t. canonical bundles in lexicon

• Lattice Rescoring
  – Rescore Switchboard ASR lattices generated by SRI
Landmark / Feature Based Model of Human Perception

Analysis-by-Synthesis:
Incorporating higher level of linguistic knowledge for re-evaluating hypothesized word sequences
[Stevens, 2000]

[From Stevens, 2002]
2. Exploiting Speech Production Models in ASR

• Statistical methods for phonological distinctive feature (PDF) detection

• Incorporating distinctive feature knowledge in ASR model structure

• **Articulatory models of vocal tract dynamics**

• Integrating distinctive features in traditional ASR systems
Articulatory Models of Vocal Tract Dynamics

Message:
Time-Aligned Phonetic Transcription

| p1 | p2 | p3 | p4 | p5 |

Phone segmentation

Target Path

Articulatory Trajectory

Acoustic Features (formants)

Bakis, 1993
Articulatory Models of Vocal Tract Dynamics

- Multi-dimensional articulatory models obtained as the Cartesian product models for each articulator dimension result in enormous computational complexity during search.

- Use traditional ASR to generate hypothesized phonetic transcriptions:

  \[\hat{H} = \arg \max_H D\left(O^H, O^T\right)\]

- Choose the phonetic transcription that is the most “plausible” according to the articulatory model.

Acoustic Features \(O^T\) → \[
\begin{array}{c}
\text{HMM Based ASR} \\
\text{Hypothesized} \\
\text{Phonetic Transcriptions}
\end{array}
\]

\[
\begin{cases}
H_1 \\
\vdots \\
H_M
\end{cases}
\]

→ Articulatory Model →

\[
\begin{cases}
O^{H_1} \\
\vdots \\
O^{H_M}
\end{cases}
\]

Generated Acoustics
Articulatory Models of Vocal Tract Dynamics

- **Coarticulation**
  - Empirically designed FIR filters [Bakis]
  - Deterministic hidden dynamic model (HDM) [Bridle et al, 1999]
  - Vocal tract resonance dynamics (VTR) [Deng et al, 1998]

- **Articulatory-to-Acoustic Mapping**
  - Radial basis functions [Bakis]
  - MLPs [Bridle et al, 1999]
2. Exploiting Speech Production Models in ASR

• Statistical methods for phonological distinctive feature (PDF) detection

• Incorporating distinctive feature knowledge in ASR model structure

• Articulatory models of vocal tract dynamics

• Integrating distinctive features in traditional ASR systems
Integrating Speech Production Models in Traditional ASR Systems

• PDF’s as features in hidden Markov model ASR

• Disambiguating HMM based ASR lattice hypotheses through PDF re-scoring

• Review of the relationship between vocal tract shape and acoustic models

• Articulatory based model normalization / adaptation
PDFs as Features in HMM-Based ASR

• **PDF Integration / Synchronization** [Kirchhoff et al, 2000] [Stuker et al, 2003][Metz et al, 2003]
  - Coupled Features – Single observation stream: \( P(s_k | X) \)
  - Independent Features – Separate streams of PDFs integrated at the state level:
    \[
    \prod_{i=1}^{N} P(s_t | X_t^i)
    \]
  - Unsynchronized Features – Use of syllable rather than phone-based acoustic units
    • Articulatory synchronization believed to occur at syllable boundaries
Disambiguating ASR Hypotheses by PDF Rescoring

- Used for re-scoring TIMIT phone lattices [Rose et al, 2006]
- PAC increase from 69.1% to 72.5% with PDF re-scoring
Confusion Network Combination

• Are different Phonological Distinctive Feature systems complementary?
• Combine phone lattices from features obtained from 3 different systems:
  – Multi-valued features (MV)
  – “Sound Patterns of English” features (SPE)
  – Government Phonology (GP)

Phonological Distinctive Feature Vectors
Phonological Lattices

Confusion Network Combination And Re-Score
Consensus String

ASR

MFCC

MV PDF Detector
SPE PDF Detector
GP PDF Detector

ASR
ASR
ASR
Confusion Network Combination

• Combine phone lattices produced from multiple DFDs …

… Into a confusion network …

… and re-score

<table>
<thead>
<tr>
<th>TIMIT Phone Recognition Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
</tr>
<tr>
<td>MFCC+GP+MV+SPE</td>
</tr>
</tbody>
</table>
Integrating Speech Production Models in Traditional ASR Systems

- PDF’s as features in hidden Markov model ASR

- Disambiguating HMM based ASR lattice hypotheses through PDF re-scoring

- **Review of the relationship between vocal tract shape and acoustic models**

- Articulatory based model normalization / adaptation
Speech Production Model for Voiced Sounds

\[ \sum_{n=\infty}^{\infty} u_o(t-nT) = \text{Glottal Pulse} \]

\[ G_p(s) \rightarrow H(s) \rightarrow R(s) \rightarrow s(t) \]

Relate sound pressure level at the mouth, \( s(t) \), to the volume velocity at the glottis, \( u(t) \)

\[ V(s) = G_p(s)H(s)R(s) \]

Glottal Pulses: Input Volume Velocity

Sound Pressure Level at the Mouth
Vocal Tract Model

Model assumptions:
- Quasi-steady flow from pulsating jet in the larynx (more on this latter)
- Plane wave propagation through a series of concatenated acoustic tubes (cross sectional area << wave length)

Typical Wavelength: \[ \lambda = \frac{c}{f} \approx \frac{331 \text{ m/s}}{100 \text{ Hz}} = 3.3 \text{ meters} \]

Typical Cross Sectional Area: \[ \text{Area} \approx 3 \text{ cm} \]

Vocal Tract Shape \quad \rightarrow \quad \text{Formants}

1. Wave equation for acoustic tube
2. Acoustic tube transfer function
3. Tube formants
From Vocal Tract Shape to Formants – Acoustic Tube Model


Cylindrical Tube of length $dx$:

- Motion of Air through tube is characterized entirely by
  - Volume velocity: $u(x,t) = U(x)e^{st}$
  - Pressure: $p(x,t) = P(x)e^{st}$

\[
\begin{align*}
A & \quad \text{Cross sectional area} \\
\rho & \quad \text{Density of air} \\
\rho Adx & \quad \text{Mass of air in tube} \\
P_o & \quad \text{Atmospheric pressure} \\
P_o + p(x,t) & \quad \text{Total pressure in tube}
\end{align*}
\]
The relationship between current and voltage in the electrical circuit is equivalent to the relationship between volume velocity and pressure in the acoustic tube.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Acoustic</th>
<th>Electrical</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x,t)$</td>
<td>Pressure</td>
<td>Voltage</td>
</tr>
<tr>
<td>$u(x,t)$</td>
<td>Volume Velocity</td>
<td>Current</td>
</tr>
<tr>
<td>$L = \rho / A$</td>
<td>Inertance</td>
<td>Inductance</td>
</tr>
<tr>
<td>$C = A / \rho c^2$</td>
<td>Compliance</td>
<td>Capacitance</td>
</tr>
<tr>
<td>$R$</td>
<td>Viscous Friction</td>
<td>Series Resistance</td>
</tr>
<tr>
<td>$G$</td>
<td>Heat Loss</td>
<td>Shunt Conductance</td>
</tr>
</tbody>
</table>
Electrical Analog of Acoustic Tube

Apply Kirchoff’s Laws to get:

1. Coupled Wave Equations:

\[
\frac{dP(x)}{dx} = -zU(x)
\]

\[
\frac{dU(x)}{dx} = -yP(x)
\]

where: \( z = Ls + R \) \( y = Cs + G \)

2. Time Independent Wave Equations:

\[
\frac{d^2 P(x)}{dx^2} = zyP(x)
\]

\[
\frac{d^2 U(x)}{dx^2} = zyU(x)
\]
Find Transfer Function of a Single Acoustic Tube

**Glottis:**
Acoustic: closed ended
Electrical: open circuit

**Lips:**
Acoustic: open ended
Electrical: short circuit

Transfer Function

\[
H(s) = \frac{U_\ell}{U_g} = \frac{U(0)}{U(-\ell)}
\]

Solution to Coupled Wave Equations:

\[
U(x) = U_+ e^{\gamma x} + U_- e^{-\gamma x} \\
P(x) = P_+ e^{\gamma x} + P_- e^{-\gamma x}
\]

where propagation constant is: \( \gamma = \pm \sqrt{zy} \)

Transfer Function:

\[
H(s) = \frac{U(0)}{U(-\ell)} = \frac{1}{\cosh \gamma \ell}
\]
Acoustic Tube Resonant Frequencies

Poles of Transfer Function: \(H(s) = \frac{1}{\cosh \gamma \ell}\)

for the lossless case \((R=G=0)\): \(\gamma = \left[(sL + R)(sC + G)\right]^{\frac{1}{2}} = j\omega \sqrt{LC}\)

occur when: \(\omega_n \sqrt{LC} \ell = \frac{(2n-1)}{2}\pi\)

\[\Rightarrow f_n = \frac{1}{4\sqrt{LC} \ell} (2n-1)\]

Typical Values: \(\ell = 17.5cm\) \(\sqrt{LC} = \sqrt{\frac{\rho A}{c^2}} = \frac{1}{c} \approx 0.003\) \(\Rightarrow f_1 \approx 500Hz\)

Transfer function for lossless acoustic tube contains equally spaced, zero bandwidth spectral resonances (formants):
Frequency Warping Based Speaker Normalization

- Single tube model of reduced shwa vowel with length 17.5 cm will have formant frequencies 500 Hz, 1500 Hz, 2500 Hz, ...

- Tube length $\ell$ and formant frequencies will vary among speakers according to $f_n \approx (2n - 1) / 4\ell c$

- Implies that the effects of speaker dependent variability can be reduced by frequency normalization
Frequency Warping Based Speaker Normalization

- Normalize for speaker specific variability by linearly warping frequency axis, $f' = \alpha f$
- Warping can be performed by warping the mel-scale filter-bank [Lee and Rose, 1998]

$$\alpha_i Y_i = \alpha_i O_i$$

- Optimum warping factor found by performing ensemble search to maximize $P(O_\alpha | \lambda)$

$$\tilde{\alpha} = \arg\max_\alpha P(O_\alpha | \lambda)$$

- HMM model is trained from warped utterances to obtain a more “compact” model
Relationship Between Vocal Tract Shape and Formants

- In general, formant frequencies for different phonemes are a more complicated function of vocal tract shape:

  ![Diagram of American English Vowels](image)

  [Jurafsky and Martin, 2008]

- **Suggests that frequency warping based speaker normalization should be phoneme or PDF dependent …**
Time Dependent Frequency Warping Based Speaker Normalization

- Localized estimates of frequency warping based speaker normalization transformations can be obtained by optimizing a global criterion.

- Implement a decoder that simultaneously optimizes frame based acoustic likelihood and warping likelihood.

- **Augment the state space of the Viterbi decoder in ASR** [Miguel et al, 2005]

- There must be other speech production oriented adaptation normalization approaches!
Augmented State Space Acoustic Decoder

- “3D” Trellis: Augment HMM state space to incorporate warping factor ensemble [Miguel et al, 2008]

\[
\phi_{j,n}(t) = \max_{i \in I, \alpha_m \in A} \{ \phi_{j,m}(t-1) a_{i,j}^{m,n} b_j(c_t^{\alpha_m}) \}
\]

• Modified Viterbi Algorithm:

Decoding and Normalization Performed in a Single Pass
Frequency Warping Based Speaker Normalization

- Modify frequency warping based normalization to facilitate global optimization of frame based frequency warping

Utterance of the word “two”

Frame based Warping function likelihoods

- Augmented state space decoder – ML procedure to select from a discrete ensemble of warping functions for each frame

[Miguel et al, 2005]
3. Resources

• Articulatory Measurement and Clinical Tools

• Corpora

• Workshops
Direct Articulatory Measurements

3D Articulagraph in Edinburgough Speech Production Facility

2D EMA Trajectories from Oxford University Phonetics Lab

Electropalatograph (EPG) from UCLA Phonetics Lab

Linguopalatal contact measurements for different prosodic positions
“Partial” Direct Measurements - Visual Information

- Partial direct articulatory measurements fused with acoustic information in audio-visual ASR [Potamianos et al, 2004]
“Partial” Direct Measurements – Glottal Information

- **Glottal Electro-Magnetic Sensors (GEMS):**
  - Very low power radar-like sensors [Burnett et al, 1999]
  - Positioned Near Glottis: Measures motion of rear tracheal wall
  - Developed at Laurence Livermore and Commercialized by Aliph

- Research programs have investigated their use in very high noise environments
Hot-Wire Anemometer and Vocal Tract Aerodynamics

• Hot-Wire Anemometers have been used for verifying aeroacoustic models of phonation [Mongeau, 1997]

Apparatus for simulating the excitation of plane waves in tubes by small pulsating jets through time varying orifices [Mongeau, 1997]
Clinical Tools - MRI and EEG

EEG Sensors in McGill Speech Motor Control Lab

Averaging of signals to separate evoked responses to various stimuli from background activity

Magnetic Resonance Imaging in McGill Speech Motor Control Lab

MRI images – Relationship between perception and articulatory motor control [Pulvermuller, 2006]
Resources – Corpora

• Phonetically labeled speech corpora
  – TIMIT
  – ICSI Switchboard transcription project [Greenberg, 2000]
  – Buckeye Corpus (Ohio State)
  – Switchboard [King et al, 2006]

• Direct Articulatory Measurements
  – Wisconsin x-ray microbeam articulatory corpus
  – MOCHA – Parallel acoustic articulatory recordings (EMA, EPG, EGG measurements) of a handful of speakers reading ~450 sentences (Edinburgh) [Wrench et al, 2000]
  – Audio-Visual TIMIT corpus (AVTIMIT) [MIT]
  – CUAVE – Audio-visual corpus [Patterson, 2002]
Resources – Workshops

• **U.S. Government Sponsored JHU Workshops**
  – 1997 – Doddington et al – Syllable-based speech processing
  – 1998 – Bridle et al – Segmental hidden dynamical models for ASR
  – 2004 – Hasagawa-Johnson et al – Landmark based speech recognition
  – 2006 – Livescu et al – Articulatory feature based speech recognition
Speech Production Topics Not Covered

• **Manifold Based Approaches**
  – Assume that speech itself is constrained to lie in some subspace but we don not know the dimensionality of the subspace
  – Laplacian Eigenmaps, Locality Preserving Projections, ISOMAP
  – Consider practical gains from mapping data onto a space of intrinsic dimension associated with a non-linear manifold [He and Niyogi][Nilson and Kleijn][Tang and Rose]

• **Speech modeling based on nonlinear vocal tract airflow dynamics** [Maragos et al]