Tutorial 6 - T6
Sunday Afternoon, Sep 2, 2:00PM - 5:30PM

Articulatory Representations: Measurement, Estimation, and Application to State-of-the-Art Automatic Speech Recognition

Carol Espy-Wilson - University of Maryland, USA
Mark Tiede - Haskins Lab, USA
Hosung Nam - Korea University, Seoul
Vikramjit Mitra - Apple Inc., USA
Ganesh Sivaraman - Pindrop, Atlanta, USA
Articulatory representations have been studied and applied to various speech technologies for many years. One of the persisting challenges of articulatory research has been the paucity of reliable articulatory data and non-scalability to large scale subject-independent applications. The aim of this tutorial is to present an overview of research in articulatory representations for large scale applications. This tutorial will discuss best practices in articulatory data collection and synthesis, present recent developments in subject independent acoustic-to-articulatory speech inversion and describe state-of-the-art Convolutional Neural Network (CNN) architectures for large vocabulary continuous speech recognition incorporating both articulatory and acoustic features. This tutorial combines experts from scientific and engineering backgrounds to present a concise tutorial about articulatory features, their measurement, synthesis and application to state-of-the-art Automatic Speech Recognition.

Carol Espy-Wilson, the lead organizer is a Professor in the Department of Electrical and Computer Engineering and the Institutes for Systems Research at the University of Maryland, College Park. She has more than 3 decades of leading research endeavors in fundamental speech acoustics, vocal tract modeling, speech and speaker recognition, speech segregation, speech enhancement, and more recently speech inversion. One of the tools coming out of Dr. Espy-Wilson's lab is a vocal tract modeling tool, VTAR that many scientists and engineers have downloaded to use for research and teaching. Dr. Espy-Wilson has advised a number of PhD and MS students and published many papers in reputed academic conferences and journals.

Mark Tiede is a Senior Scientist at Haskins Laboratories active in speech production research, particularly the study of dyadic interaction. He has more than 20 years experience in the use of point source tracking methods such as electromagnetic articulography (EMA) and is the author of a widely used tool for the use of analyzing such data (mvview).

Hosung Nam is an Assistant Professor in the Department of English Language and Literature at Korea University, Seoul, South Korea. He received the M.S. and Ph.D. degrees from the Department of Linguistics at Yale University, New Haven, CT, in 2007. He is a linguist who is an expert in the field of articulatory phonology. His research emphasis is on the link between speech perception and production, speech error, automatic speech recognition, sign language, phonological development, and their computational modeling. He has been a Research Scientist at Haskins Laboratories, New Haven, since 2007.

Vikramjit Mitra is a Research Scientist at Apple Inc. He received his Ph.D. in Electrical Engineering from University of Maryland, College Park; M.S. in Electrical Engineering from University of Denver, B.E. in Electrical Engineering from Jadavpur University, India. His research focuses on signal processing for noise/channel/reverberation, speech recognition, production/perception-motivated signal processing, information retrieval, machine learning and speech analytics. He is a senior member of the IEEE, an affiliate member of the SLTC and has served on NSF panels.

Ganesh Sivaraman is a Research Scientist at Pindrop. He received his M.S. and Ph.D. in Electrical Engineering from University of Maryland College Park, B.E. in Electrical Engineering from Birla Institute of Technology and Science, Pilani, India. His research focuses on speaker independent acoustic-to-articulatory inversion, speaker adaptation, speech enhancement and robust speech recognition.
Articulatory Representations: Measurement, Estimation, and Application to state-of-the-art Automatic Speech Recognition

- **Prof. Carol Espy-Wilson** – “Acoustic Phonetics, Speech Acoustics and Vocal Tract Modeling”  
  Dept. Electrical and Computer Eng. & Institute for Systems Research, University of Maryland College Park, USA

- **Dr. Mark Tiede** – “Observing and Measuring Speech Articulation”  
  Haskins Laboratories, USA

- **Dr. Ganesh Sivaraman** – “Estimation of articulatory representations from acoustics”  
  Pindrop, Atlanta, USA

- **Dr. Vikramjit Mitra** – “Application to state-of-the-art Automatic Speech Recognition”  
  Apple Inc., USA  
  Affiliate, University of Maryland College Park

Acoustic Phonetics, Speech Acoustics and Vocal Tract Modeling

Carol Espy Wilson  
Dept. Electrical and Computer Eng. & Institute for Systems Research, University of Maryland
Speech Chain (Denes and Pinson, 1972)

Articulation and Acoustics

- **Phonetics**
  - Articulatory Phonetics: how speech sounds are produced
  - Acoustic Phonetics: the transmission and physical properties of the speech signal
  - Auditory Phonetics: perception of speech sounds

- **Speech Acoustics**: relationship between vocal tract shape (glottis to lips) and acoustic properties of the speech signal
Mid-Sagittal View of Human Vocal Tract (Parsons, 1986)

Source-Filter View of Speech Production (Fant, 1960; Stevens, 1998)

Modulated airflow through glottis

Vocal tract (glottis to lips)

Radiation of sound

Listener’s ear, microphone

e(t) \rightarrow v(t) \rightarrow r(t) \rightarrow s(t)

\begin{align*}
E(\omega) & \quad V(\omega) & \quad R(\omega) & \quad S(\omega) \\
\text{s(t)} & = e(t) * v(t) * r(t) \\
S(\omega) & = E(\omega) V(\omega) R(\omega)
\end{align*}
Laryngeal High-Speech Videoendoscopy (Mehta et al., 2011)

\[ e(t) \rightarrow v(t) \rightarrow r(t) \rightarrow s(t) \]
Laryngeal High-Speech Videoendoscopy (Mehta et al., 2011)

\[ e(t) \rightarrow v(t) \rightarrow r(t) \rightarrow s(t) \]

Model of Glottal Source (Peterson & Barney, 1952; Stevens, 1998)

\[ F_o = \frac{1}{2\pi} \sqrt{\frac{1}{MC}} \]

\( M \) = mass of the vocal folds
\( C \) = compliance of the vocal folds
Glottal Source

(Sadaoki Furui, 1989)

Voice Quality (Klatt & Klatt, 1990)
Voice Quality (Mehta et al., 2011)

(creaky) (modal) (breathy)

Aspiration Source (Stevens 1999)

Rapid air flow through the glottal constriction that forms a jet which impinges on the walls of the vocal tract above the glottis
Frication Source (Stevens 1999, Shadle 1985)

Rapid airflow through a narrow constriction that forms a jet which impinges on an obstacle downstream from the constriction.

Vocal Tract Transfer Function
Source-Filter View of Speech Production (Stevens 1998)

Source Spectrum
Vocal tract transfer function
Radiation Characteristics
Power spectrum of speech signal

Framework: Phonetic Feature Theory (Chomsky & Halle, 1968)

- 20 or so phonetic features characterizing all of the world’s languages
- Based on the position of the articulators
- Minimal unit that distinguishes between sounds

zip + voiced
sip - voiced
Differ in Source Phonetic Feature

tip - continuant
sip + continuant
Differ in Manner Phonetic Feature

ship - anterior
sip + anterior
Differ in Place Phonetic Feature
Source Phonetic Features

Initial sound in “sip” -voiced
Initial sound in “zip” +voiced

Source feature voiced

lack of striations -voiced
vertical striations +voiced
Manner Phonetic Features

No constrictions  
Vowel “app”  
+ sonorant  
+ syllabic

Narrow constrictions  
Fricative “sap”  
- sonorant  
- syllabic  
+ continuant

Complete closure  
Stop “tap”  
- sonorant  
- syllabic  
- continuant

Velum lowered  
Nasal “nap”  
+ sonorant

Manner feature sonorant

Energy concentrated at high frequencies  
“Sprouted”  
-sonorant

Energy concentrated at low frequencies  
+sonorant
Place Phonetic Feature

+ labial /b/ and /p/
+ alveolar /d/ and /t/
+ velar /g/ and /k/
+ labiodental /v/ and /f/
+ dental /dh/ and /th/
+ alveolar /z/ and /s/
+ postaveolar /zh/ and /sh/

Place Feature *Labial vs. Alveolar*

spectral prominence

falling

rising

labial /p/
alveolar /t/
Vocal Tract Transfer Function

\[ e(t) \xrightarrow{v(t)} r(t) \xrightarrow{s(t)} \]

Peterson and Barney Vowel Data (1952)
Phonetic Feature Hierarchy

- sonorant
  - syllabic
    - continuant
      - nasal
        - labial
        - alveolar
        - velar
      - lateral
      - palatal
      - labial w
      - palatal y
    - rhotic
      - labial
      - alveolar
      - palatal
      - dental
      - labial v
      - alveolar n
      - labial m
      - velar ng

- strident
  - voiced
    - labial
    - alveolar
    - palatal
    - dental
    - labial v
    - alveolar n
    - labial m
    - velar ng

Place Feature Alveolar (/s/) vs. Alveolo-Palatal (/sh/) for Fricatives

"seashore"

kHz

Time (sec)
Importance of Relative Measures to Reduce Speaker Characteristics (Bitar & Espy-Wilson, 1995)

![Graphs showing energy levels for different frequency bands for males and females.]

Phonetic Feature Parameters vs. MFCCs (Deshmukh et al., 2002; Pruthi & Espy-Wilson, 2004)

Task: Digit Database, Adult: TI-46, Children: TIDIGIT

<table>
<thead>
<tr>
<th>Train</th>
<th>Adult</th>
<th>Female</th>
<th>Male</th>
<th>Adult</th>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>Adult</td>
<td>Male</td>
<td>Female</td>
<td>Child</td>
<td>Adult</td>
</tr>
<tr>
<td>MFCCs (39 pars)</td>
<td>99.88</td>
<td>68.29</td>
<td>70.27</td>
<td>60.20</td>
<td>62.37</td>
</tr>
<tr>
<td>APs (28 pars)</td>
<td>97.26</td>
<td>80.34</td>
<td>80.81</td>
<td>85.15</td>
<td>85.88</td>
</tr>
<tr>
<td>APs (30 pars)</td>
<td>99.53</td>
<td>79.24</td>
<td>90.90</td>
<td>85.70</td>
<td>89.81</td>
</tr>
</tbody>
</table>
Two Subjects (Zhou et al., 2008)

Retroflex /r/ (S1)

- Tip up
- Dorsum down

Bunched /r/ (S2)

- Tip down
- Dorsum up

Vocal Tract Shape for AE /r/

The Three Constrictions

- lips
- palate
- pharynx

And Large Volume behind Lip Constriction
3D reconstruction of vocal tract

Vocal Tract Models (Zhou et al., 2008)

- MR images
- 3-D shape
- Area function
- Simple-tube model
Acoustic Responses of Vocal Tract Models (Zhou et al., 2008)

Spectra of natural sound

3-D FEM

Area function

Simple-tube model

Four element tube model for a retroflex /r/ (Espy-Wilson et al., 2000)
F3, F4 and F5 are half-wavelength resonances of the cavity posterior to the palatal constriction fairly evenly spaced.

Half-wavelength tube:

\[ f_n = \frac{c}{2L_b} n \]

Spacing between formants:

\[ \frac{c}{2L_5} = \frac{c}{2(12)} = 1460 \text{Hz} \]

Different Trajectories for F4 and F5

“warav”
Acoustic Signatures for Tongue Shapes used to Produce AE /r/ (Zhou et al., 2008)

Midsagittal MR slices

Spectra of sustained sound

Spectrogram of ‘warav’

Lenition: undershoot in articulation (Espy-Wilson, 1992)
Differences in Articulatory Strategies (Deshmukh et al., 2005)

Aperiodic vs. Periodic Energy (Deshmukh et al., 2005)
Results on Manner Classes (Deshmukh et al., 2005)

<table>
<thead>
<tr>
<th></th>
<th>Only strong periodic energy</th>
<th>Only strong aperiodic energy</th>
<th>Strong periodic and aperiodic energies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sonorants</td>
<td>88.5</td>
<td>5.4</td>
<td>6.1</td>
</tr>
<tr>
<td><strong>Voiced Obstruents</strong></td>
<td><strong>33.29</strong></td>
<td><strong>44.62</strong></td>
<td><strong>22.09</strong></td>
</tr>
<tr>
<td>Unvoiced Obstruents</td>
<td>2.2</td>
<td>97.6</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Different F3 trajectories in “Cart”
F3 and Tongue Tip Displacement (Boyce and Espy-Wilson, 1997)

F3 trajectories across contexts (warav, wavrav, wabrav, wagrav, wadrav)
Explanation of F3 trajectories

Overlapping Gesture Hypothesis of Coarticulation (Boyce & Espy-Wilson, 1997)
Different F3 trajectories in “Cart” (coarticulation)

Overlapping gestures in fast-spoken utterance results in lower spectral peak
Lenition: undershoot in articulation  (Espy-Wilson, 1992)

Note: Due to rate of change of speech, the degree of overlap between the gestures is altered and there may be undershoot in the gestural target, but the overall gestural pattern is similar.
References


References


Observing and Measuring Speech Articulation

mark tiede
Overview

- Motivation: why study speech articulation?
  - speech acoustics are a surface property of the underlying movements of vocal tract articulators
  - learning about how these movements are coordinated in time and space informs us on how:
    - some output sounds are not possible, while others are achievable using different vocal tract configurations (e.g. /u/)
    - some patterns of movement result in phonetic lenition, reduction, and sandhi (which must otherwise be stipulated)
    - coproduction of speech targets mutually influences articulator movements and their acoustics (coarticulation)
    - relative timing of movements characterizes rate and style effects (e.g. 'casual' vs. 'formal'); dialect; individual differences; etc.

- Methods for observing speech articulation
  - Planning (e.g. fMRI, EEG, ECoG)
  - Activation (e.g. EMG, plethysmography)
  - Execution: point source methods (e.g. EMA, MoCap); full profile methods (e.g. rtMRI, ultrasound, cineradiography)

- Quantifying speech articulation
  - practical considerations for converting to a noise-corrected, head-centric coordinate system
  - heuristics for evaluating the extent and relative phasing of speech articulator movements

- Applications of speech articulatory data
  - studying synergies among speech articulators (Task Dynamics, Articulatory Phonology, TaDA)
  - through the use of an estimated articulatory substrate, improvements in ASR results

Definitions

**What is speech?**

\[ \rightarrow \text{a re-purposing of oral anatomy for communication} \]

**Source Filter Theory** (Fant, 1960)

- air pressure from lungs induces vibration in vocal folds at a fundamental frequency \( F_0 \) (Source)
- shape of vocal tract (position of tongue, lips, etc.) strengthens or weakens harmonics of \( F_0 \) (Filter)
Definitions

/i/ “heed”  /u/ “who’d”  /ɑ/ “hod”

![Structural MRI of sustained English vowels](image)

Definitions

But speech is not static:
- vowels produced with dynamically varying pitch (F0), intensity, formant structure
- consonants are superimposed vocal tract constrictions
- vocal tract filter shaped by overlapping movements or gestures of the speech articulators (tongue, jaw, velum, lips) interacting with glottal source
- interaction among gestures with competing targets/goals leads results in acoustic variability (coarticulation)

Movements of the speech articulators unfold at
- potentially different, relatively slow rates
- overlap in time
- result in nonlinear, relatively fast changes in acoustic output

How can we observe these movements?
Instrumental techniques for observing speech production

**Cerebral Activity** (intention to speak)
- blood flow (hemodynamic response) methods: fMRI, fNIRS
- electrical activity: EEG, MEG, ECoG

**Motor Commands** (muscle activation)
- electromyography (EMG)

**Aerodynamics** (airflow & pressure)
- plethysmography, nasalance, Rothenburg mask

**Structural Movement** (articulators in motion)
- video methods, palatal contact (EPG)
- point tracking (EMA, XRMB, MoCap)
- full profile methods (ultrasound, rMRI, cineradiography)

**Acoustic Recordings** (vocal tract filter output)

---

**Example: electrocorticography (ECoG)**

Neural activity (gamma-band power) in the ventral sensory-motor cortex (vSMC) is correlated with measured articulator movement kinematics (position, velocity) during the production of English vowels (Conant et al., 2018, p2958)

- **pro**: very high quality cortical data
- **con**: highly invasive

Placement of ECoG grid

/a/ vowel: lips (video) and ultrasound

vSMC electrode articulator encoding
Example: muscle activation

Electromyography (EMG): measures electrical potential generated by muscle cells when they contract

e.g.: **masseter** (jaw elevation) and **ABD** (jaw lowering) response to **unexpected jaw perturbation**, collected using surface electrodes (Tiede et al., 2006)

- perturbations applied by robot coupled to jaw in random sequence to 20% of trials of an alternating vowel sequence: half **upwards**, half **downwards**

---

Example: muscle activation

Results:

- F1 maintained consistently (Panel 1) despite jaw perturbation (Panel 2)
- increased masseter (jaw elevator) when jaw perturbed **down**
- increased ABD (jaw depressor) when jaw perturbed **up**

**pro**: important for understanding timecourse of articulation planning

**con**: difficult to obtain signal from specific muscles

- /a/- /i/- /ɛ/- /i/- /a/-

---

// Interspeech-2018  Hyderabad, India  Page 33 of 103 //
**Example: palatal contact (EPG)**

**Electropalatography** measures contact between the tongue and a custom-fit palatal prosthesis.

The grid at right shows the contact pattern at the cursor (showing both /k/ and /t/ constrictions active).

**con:** prosthesis interferes with production.

---

**Example: aerodynamics**

**Pneumotachograph** (Rothenburg Mask) measures airflow separately for nasal and oral outflow.

**con:** mask interferes with production.
Example: cineradiography

Cineradiography (X-ray movie)
Stevens & Öhman (1963)

<table>
<thead>
<tr>
<th>hape</th>
<th>hatè</th>
<th>hakè</th>
</tr>
</thead>
<tbody>
<tr>
<td>hopè</td>
<td>hata</td>
<td>haka</td>
</tr>
<tr>
<td>hadè</td>
<td>hane</td>
<td>hase</td>
</tr>
<tr>
<td>hazè</td>
<td>hasa</td>
<td>hap</td>
</tr>
<tr>
<td>hét</td>
<td>hek</td>
<td>hap</td>
</tr>
<tr>
<td>hat</td>
<td>hak</td>
<td>hati</td>
</tr>
<tr>
<td>hatè</td>
<td>hatèz</td>
<td>hato</td>
</tr>
<tr>
<td>hatu</td>
<td>hatzt</td>
<td>hatat</td>
</tr>
<tr>
<td>hatab</td>
<td>hapen</td>
<td>hili</td>
</tr>
<tr>
<td>hida</td>
<td>higi</td>
<td>higa</td>
</tr>
<tr>
<td>ia</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Why did Ken set the soggy net on top of his dock?
I have put blood on her two clean yellow shoes.
Stop.

**con:** high radiation hazard

Example: ultrasound

**pro:**
- high frame rate (>100 Hz)
- non-invasive, easy to use

**con:**
- transduction stops at air-tissue boundary (thus tongue tip is rarely seen)
- images are relative to probe, which may not be accurately aligned with head
- images can be noisy, and extracting contours is not straightforward

American English vowel sequence

[Image of ultrasound scan with labels anterior and posterior]
**Example: rtMRI**

Courtesy P. Hoole, E. Kunay, C. Carignan (IPS Munich) and J. Frahm, A. Joseph (Max-Planck-Institute for Biophysical Chemistry, Göttingen)

**pro:**
- good temporal (>50 Hz) and excellent spatial resolution
- entire vocal tract imaged

**con:**
- acoustically noisy
- requires access to specialized, expensive facilities
- tracking vocal tract features is not straightforward

“Bis er bunte erklärt”

**Example: point tracking**

**X-ray Microbeam**

Fujimura et al. (1973) (U. Tokyo)

- rasterized X-ray beam active only in vicinity of pellet, thus mimimizing radiation exposure
- pellets track tongue, jaw, lips
- /a/- /i/- /a/ sequence

- **pro:** excellent temporal resolution
- **con:** radiation hazard, sparse representation of VT
Example: point tracking

X-ray Microbeam
Fujimura et al. (1973) (U. Tokyo)

same /a-i-a/ sequence showing temporal trajectories of pellets (half speed)

Example: point tracking

U.Wisc. X-ray Microbeam corpus (XRMB)
(Westbury, 1994)

- 57 speakers of American English (largest available corpus of articulatory data)
- 8 pellets track tongue, jaw, lips
- palate and pharyngeal wall traces provide estimate of VT size
Example: point tracking

**Electro-Magnetic Articulometry (EMA)**

**pro:**
- high sampling rate (>200 Hz)
- good spatial accuracy (~0.2mm)
- modern systems track three spatial and two angular orientation dimensions per sensor
- minimal radiation exposure (kHz RF)

**con:**
- limited to anterior vocal tract
- sensor coils and their wires are invasive
- sensitive to magnetic field distortions (metal)

---

**EMA: commercialization**

**AG500/501**
Carstens Medizinelectronik

**WAVE**
Northern Digital Instruments

Evaluations for speech research: Yunusova et al. (2009), Berry (2011), Savariaux et al. (2017)
EMA: how it works

Receiver Coil (sensor)

Faraday’s Induction Principle

oscillating magnetic field

induces current in core and in wire

transduced signal proportional to cube of distance from transmitter

EMA: how it works

• transmitters 1 and 2 induce voltages in Receiver Coil proportional to cube of its distance from each

• position lies at intersection of distance-determined circles centered on transmitters
EMA: how it works

**Problem:** if a Receiver Coil is not aligned with the transmitter field, the transduced signal strength is reduced by a factor equal to the cosine of the tilt angle, overestimating the distance from the transmitter

**Solution:** use three transmitters – comparison of known transmitter circle radius with transduced radius gives tilt correction factor (Perkell et al., 1992)

EMA: movement correction

EMA data is collected in a coordinate system relative to the device.
To be useful each frame of data must be corrected to references placed on the head and aligned with the spkr’s occlusal plane.
This illustration shows a CD rotated in a jewel case, with and without movement correction.
Point source data: delimiting articulatory gestures

"banana" /b/ /ah/ /n/ /ae/ /n/ /ah/

Point source data: delimiting articulatory gestures

first /n/ of "banana"

1) velocity zero identifies maximum constriction (MAXC)
Point source data: delimiting articulatory gestures

first /n/ of “banana”

1) velocity zero identifies maximum constriction (MAXC)
2) bracking velocity peaks (PVEL) determine amplitude range

Point source data: delimiting articulatory gestures

first /n/ of “banana”

1) velocity zero identifies maximum constriction (MAXC)
2) bracking velocity peaks (PVEL) determine amplitude range
3) 20% of amp. range determines gestural on/offset (GONS, GOFFS)
Point source data: delimiting articulatory gestures

1) velocity zero identifies maximum constriction (MAXC)
2) bracking velocity peaks (PVEL) determine vel. amp. range
3) 20% of amp. range determines gestural on/offset (GONS, GOFFS)

→ delimited gesture

Example: documenting endangered language contrasts

Four-way coronal stop distinction in Wubuy (Best et al. 2010)

EMA used to identify trajectories into contrastive voiced coronal constrictions

/aDa/

*ada* (alveolar)

*ada* (dental)

*ađa* (retroflex)

*ad’a* (laminal)
Example: acoustic reduction as a consequence of gestural overlap

U. Tokyo X-ray microbeam corpus (Fujimura et al., 1973)

In List: gestural overlap by /l/ masks release of /k/

In Phrase: gestural overlap by /m/ also masks release of /l/

Example: acoustic reduction as a consequence of gestural overlap

“perfec(t) memory” example used to support

- “Iceberg” Hypothesis (Fujimura 1981)
  - articulatory gestures producing consonants are highly consistent
  - they may ‘float’ relative to other icebergs to produce changes in utterance timing, but do not vary in magnitude or duration

- Articulatory Phonology (Browman & Goldstein 1990)
  - as speech rate increases gestures show increasing temporal overlap
  - apparent C deletions observed in fluent speech explainable as articulator movements whose acoustics consequences are masked by coproduced gestures, but the gestures themselves are still present underlyingly
Tiede et al. (2001):
- 8 speakers produced “perfect memory” in a carrier under three production rates
- relative phasing between /k/ and /t/ preserved across rates
- acoustic release increasingly masked as rate increased
- apical gesture preserved even when release masked

Example: acoustic reduction as a consequence of gestural overlap

Kinematic data of production rate contrasts remains sparse
- we have recently newly recorded with EMA the 720 phonetically balanced Harvard sentences (IEEE 1969) from 8 speakers at two production rates (normal, fast); Tiede et al. (2017a)
- the corpus is freely available for download for research purposes from
- this work supports ongoing improvements to a model of speech recognition featuring a Tract Variable substrate (Nam et al., 2012, Mitra et al. 2017)
Example: L2 speakers are sensitive to gestural timing differences

**Phonetic Convergence** has been observed in the kinematics of L2 speakers of American English (AE) imitating productions of a native AE partner (Tiede et al., 2017b)

- age- and gender-matched dyads paired **L1** speakers of AE with native Spanish speakers who had AE as their **L2**
- L1 and L2 speakers showed systematic differences in gestural phasing and duration between their initiated (non-imitated) productions of an AE test phrase
- L2 speakers systematically altered their gestural phasing and timing in the direction of their L1 partner when imitating the test phrase
- perceptual judgements of L2 production “nativeness” by L1 AE listeners favored imitated productions

---

Example: L2 speakers are sensitive to gestural timing differences

**Dual EMA recording** Tiede et al. (2010)

- simultaneous recording from speakers seated face-to-face using NDI WAVE and Carstens AG500 Electromagnetic Articulometer (EMA) devices
- common alignment for data from each system achieved through cross-correlation of concurrently recorded audio
Example: L2 speakers are sensitive to gestural timing differences

“She asked me to bag it in green”

L1 speaker initiates,
L2 speaker imitates

initiator and imitator roles were balanced but alternated at random

Example: L2 speakers are sensitive to gestural timing differences

“She asked me to pick it in green”

Gestural events identified on EMA sensor trajectory extrema

- TT: tongue tip (/f/, /s/)
- TD: tongue dorsum (/k/)
- LA: lip aperture (/m/)

**Duration**: /m/ - /f/

**Phasing**: (event offset as percent of duration): e.g. (/k/ - /f/) / (/m/ - /f/)
Example: L2 speakers are sensitive to gestural timing differences

\[ /s/ \text{ as } \% \text{ of } /f/ \text{ to } /m/ \text{ duration in "she asked me"} \]

- /s/ occurs systematically earlier when \textit{initiated} by L2 speakers
- phasing of /s/ shifts in direction of partner preference for both L1 and L2 under \textit{imitation}

\[ \text{L2} \]
\[ \text{imitated} \quad ** \]
\[ \text{initiated} \]

\[ \text{L1} \]
\[ \text{imitated} \quad ** \]
\[ \text{initiated} \]

Modeling / Applications

\textbf{Articulatory data} provides insight into potentially ambiguous acoustic processes:

- the organization of the vocal tract rules out certain sounds (e.g. nasalized pharyngeal stops) thus providing potential constraints on ASR search space
- movement into and away from constriction targets is conditioned by competing demands on the articulators for coproduced sounds (coarticulation), and this provides useful acoustic context for disambiguation
- such movement also constrains the “many-to-one” problem (Atal et al., 1978) associated with speech inversion
- generalized speech movements can also support improved models of Articulatory Speech Synthesis (e.g., TaDA, Nam et al., 2004), and provide insight into models of language acquisition (Kröger & Birkholz, 2007)
- speech movement patterns are individually idiosyncratic, providing a robust basis for speaker identification
- more broadly they characterize dialect, accent, and certain speech pathologies
Example: ‘many-to-one’

North American English /ʁ/ is produced with different tongue shapes

Tongue configuration types for AE /ʁ/ as identified from cineradiographs by Delattre & Freeman (1968). Adapted from Hagiwara (1985).

Broadly speaking two categories:

“bunched” (dorsal constriction)
“retroflex” (apical constriction)

Examples of corresponding AE tongue configuration types for sustained /ʁ/ (Tiede et al. 2004)

All configurations result in same rhotic acoustic signature (lowered F3), but

• transitions into and out of target are different (Boyce & Espy-Wilson, 1997)

Example: ‘many-to-one’

/copæv/  /wadæv/  /wagæv/

Coproduction in clusters results in different shapes for /r/
Example: ‘many-to-one’

/Coproduction in clusters results in different shapes for /r/

Centered on F3 minimum

Example: synthesis using TaDA

Nam et al. (2004)

**TTS rules** (derived from articulatory data) generate

**Coupling Graphs**, which estimate a

**Gestural Score** (top panel gating functions), which control

**Tract Variable** activations, which drive

**Articulator Trajectories** (bottom panels), from which a time-varying vocal tract

**Area Function** is obtained for synthesis (left bottom panel)

“O say can you see by the dawn’s early light”
Speech Acoustics reflect the underlying coordination of skilled articulatory movements within the vocal tract

- speech movement patterns are individually idiosyncratic, providing a robust basis for speaker identification
- more broadly they characterize dialect, accent, and certain speech pathologies
- understanding the synergies among articulators recruited to achieve specific speech goals has advantages for applications in both recognition (by constraining search space) and synthesis (by leveraging vocal tract organization in place of arbitrary stipulation)
- modern methods are improving access to the observation of speech articulation while reducing invasive disruption

~ 2,500 years ago the great scholar Pāṇini described and organized the acoustic sounds of language on articulatory principles (the Śivasūtras of his Aṣṭādhyaṇī)

It has been a pleasure to discuss how articulation continues to inform our ideas about language with you today – thank you for your attention!

References


Estimation of Articulatory representations from acoustics
Ganesh Sivaraman
Pindrop, Atlanta GA

Overview

- Types of articulatory features
  - Discrete phonetic features
  - Continuous articulatory movement features
- Methods of estimating discrete articulatory features
  - Annotating data with discrete features
  - Methods of inversion mapping
- Methods of estimating continuous articulatory data
  - Approaches for performing speech inversion
- Performance of speech inversion systems across variabilities
  - Speaking rate
  - Speakers
- Addressing speaker variability
  - Speaker adaptation methods for speech inversion
- Future Directions
Types of Articulatory features

- **Discrete articulatory features:**
  - Features that represent articulatory configuration through parameters that take discrete values
    - Phonetic categorical features
    - Quantized articulatory positions

- **Continuous articulatory features:**
  - Features that represent measured or simulated movement of vocal tract articulators.

Motivated by Theories of phonology

- **Generative phonology:** An underlying (phonemic) string, which is transformed via a set of rules to a surface (phonetic) string (Chomsky, Halle 1968)

- **Autosegmental phonology:**
  - Phonological representation consists of multiple strings, or tiers, corresponding to different linguistic features.
  - Motivated by the observation that some phenomena of feature spreading are more easily explained if a single feature value is allowed to span more than one segment.

- **Articulatory phonology:**
  - Speech can be decomposed into a constellation of Gestures (Browman & Goldstein 1989)
  - A gesture is a constriction action along the vocal tract
Distinctive binary feature composition (Chomsky, Halle, 1968)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Articulatory correlates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocalic consonantal</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>Anterior</td>
<td></td>
</tr>
<tr>
<td>Coronal</td>
<td></td>
</tr>
<tr>
<td>Tongue position</td>
<td></td>
</tr>
<tr>
<td>Place</td>
<td></td>
</tr>
</tbody>
</table>

Consonants place of articulations in relation to the distinctive features:

<table>
<thead>
<tr>
<th>Place</th>
<th>Labial</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Palatal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anterior</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Coronal</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

Articulatory correlates of distinctive features

Place features for vowels

<table>
<thead>
<tr>
<th>Feature</th>
<th>Articulatory correlate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Tense</td>
<td></td>
</tr>
<tr>
<td>Round</td>
<td></td>
</tr>
</tbody>
</table>
Articulatory correlates of distinctive features

Features for fricative consonants

<table>
<thead>
<tr>
<th>Feature</th>
<th>Articulatory correlate</th>
<th>v</th>
<th>f</th>
<th>th</th>
<th>z</th>
<th>s</th>
<th>sh</th>
</tr>
</thead>
<tbody>
<tr>
<td>voiced</td>
<td>Vocal vocal vibration</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>strident</td>
<td>airstream from the</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>constriction hits an</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>obstacle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alveolar</td>
<td>Tongue tip against</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>alveolar ridge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>labial</td>
<td>Constriction at lips</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Place and manner features for sonorant consonants

<table>
<thead>
<tr>
<th>Feature</th>
<th>Articulatory correlate</th>
<th>w</th>
<th>r</th>
<th>l</th>
<th>y</th>
<th>n</th>
<th>m</th>
<th>ng</th>
</tr>
</thead>
<tbody>
<tr>
<td>nasal</td>
<td>Closed oral cavity,</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td></td>
<td>flow through nasal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cavity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>labial</td>
<td>Constriction at lips</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alveolar</td>
<td>Tongue tip against</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>alveolar ridge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rhotic</td>
<td>Curled up tongue</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral</td>
<td>Lateral airflow around</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>one or both sides of</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>tongue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>round</td>
<td>Lip rounding</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multilevel discrete articulatory categories. (Kirchoff, 1999)

- Feature categories and the corresponding feature classes for American English*:

<table>
<thead>
<tr>
<th>Articulatory Feature</th>
<th>Categories/classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voicing</td>
<td>Voiced, Voiceless, Silence</td>
</tr>
<tr>
<td>Manner</td>
<td>Vowel, Nasal, Lateral, Approximant, Stop, Fricative</td>
</tr>
<tr>
<td>Place</td>
<td>Dental, Coronal, Labial, Retroflex, Velar, Glottal, High, Mid, Low, Silence</td>
</tr>
<tr>
<td>Front-back</td>
<td>Front, Back, Nil, Silence</td>
</tr>
<tr>
<td>Lip Rounding</td>
<td>+Round, –Round, Nil, Silence</td>
</tr>
</tbody>
</table>

- Example phonemes and their corresponding features:

- *Appendix A.3, K. Kirchoff, Robust speech recognition using articulatory information, PhD. Thesis, 1999
**Multilayer Perceptron as classifiers for discrete AFs**

- **Groundtruth AFs**: Rule based mappings from aligned phone transcriptions to articulatory features (as defined in the previous slide).
- **Kirchoff (1999)** applied Multilayer Perceptron (MLP) for learning the mapping from acoustics to articulations.
- 5 different MLPs. One each for – Voicing, Manner, Place, Front-back, and Lip rounding.
- **Acoustic features**:
  - 8 log-RASTA PLP coefficients (Hermansky and Morgan, 1994) for clean speech system
  - 15 Modulation Spectrogram features (Greenberg & Kingsbury, 1997; Kingsbury et al., 1998).
  - Context of 5 to 9 frames as input to the MLP
- **Single layer perceptrons** with tanh activations and sigmoid outputs

---

**Acoustic modeling using discrete AFs (Kirchoff 2002)**

- ANN/HMM system for ASR
- Shallow single hidden-layer neural network to estimate AFs from acoustic features – RASTAPLP or MODSPEC.
- AFs used as features of Hidden Markov Model based ASR system.
- Acoustic and articulatory system combinations
  - HMM State level combination
  - Word score level combination
  - Feature level combination
Phone broad classes and AFs

- **Sonorant**
  - **Syllabic**
  - **Continuant**

- **Vowels**
  - **Nasal**
  - **Rhotic**

- **Consonants**
  - **Lateral**
  - **Labial**
  - **Palatal**
  - **Velar**
  - **Labio-Dental**
  - **Labio-Alveolar**
  - **Alveolar**
  - **Dental**
  - **Palatal**
  - **Velar**

---

SVM based probabilistic framework for obtaining AFs (Juneja, Espy-Wilson, 2008)

- A probabilistic hierarchical model based on decoding a speech utterance into broad class segments and further classifying the broad class segments using place and voicing features to obtain phonemes and words.
Definitions: Pronunciation and observation modeling

- **Language model**
  \[ P(w) \]

  \[ w = \text{“makes sense...”} \]

- **Pronunciation model**
  \[ P(q|w) \]

  \[ q = \{ m m m e y 1 e y 1 e y 2 k 1 k 1 k 1 k 2 k 2 k 2 s ... \} \]

- **Observation model**
  \[ P(o|q) \]

  \[ o = \]

(Livescu et al. JHU summer workshop, 2006)
Feature set for pronunciation modeling (Livescu et al. JHU summer workshop, 2006)

- Based on articulatory phonology (Browman & Goldstein ‘90) adapted for pronunciation modeling (Livescu ‘05)

<table>
<thead>
<tr>
<th>LIP–LOC</th>
<th>Protruded, Labial, Dental</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIP–OP</td>
<td>Closed, Critical, Narrow, Wide</td>
</tr>
<tr>
<td>TT–LOC</td>
<td>Dental, Alveolar, Palato–Alveolar, Retroflex</td>
</tr>
<tr>
<td>TB–LOC</td>
<td>Palatal, Velar, Uvular, Pharyngeal</td>
</tr>
<tr>
<td>TT–OP, TB–OP</td>
<td>Closed, Critical, Narrow, Mid–Narrow, Mid, Wide</td>
</tr>
<tr>
<td>GLO</td>
<td>Closed (stop), Critical (voiced), Open (voiceless)</td>
</tr>
<tr>
<td>VEL</td>
<td>Closed (non–nasal), Open (nasal)</td>
</tr>
</tbody>
</table>

- Under some simplifying assumptions, can combine into 3 streams

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tongue (cardinality 19)</td>
<td>D–CR–U–M, A–CL–U–N, ...</td>
</tr>
<tr>
<td>Glottis/velum (cardinality 3)</td>
<td>C–VO, C–VL, O–VO</td>
</tr>
</tbody>
</table>

Pronunciation variation from articulatory perspective

<table>
<thead>
<tr>
<th>feature</th>
<th>values</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLO</td>
<td>open</td>
</tr>
<tr>
<td>VEL</td>
<td>closed</td>
</tr>
<tr>
<td>TB</td>
<td>mid / uvular</td>
</tr>
<tr>
<td>TT</td>
<td>critical / alveolar</td>
</tr>
<tr>
<td>phone</td>
<td>s</td>
</tr>
<tr>
<td>phone</td>
<td>n</td>
</tr>
<tr>
<td>phone</td>
<td>s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>feature</th>
<th>values</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLO</td>
<td>open</td>
</tr>
<tr>
<td>VEL</td>
<td>closed</td>
</tr>
<tr>
<td>TB</td>
<td>mid / uvular</td>
</tr>
<tr>
<td>TT</td>
<td>critical / alveolar</td>
</tr>
<tr>
<td>phone</td>
<td>s</td>
</tr>
<tr>
<td>phone</td>
<td>n</td>
</tr>
<tr>
<td>phone</td>
<td>t</td>
</tr>
</tbody>
</table>

(Livescu et al. JHU summer workshop, 2006)
Dynamic Bayesian Networks for Multi-stream AF-based pronunciation models

- **Phone-based**
  
  ![Phone-based Diagram]
  
  - $q$ (phonetic state)
  - $o$ (observation vector)

- **AF-based**
  
  ![AF-based Diagram]
  
  - $q_i$ (state of AF $i$)
  - $o$ (obs vector)

### Strengths

- Since AFs will generally occur in more than one phone, training data for these features can effectively be shared across phones.
- It is likely that different aspects of articulation exhibit different degrees of robustness and do not deteriorate (in terms of their ability of being recognized correctly) to the same degree under adverse acoustic conditions. (Kirchoff 1999)

### Weaknesses

- Derived from phonetic transcriptions and alignments
  - Which can be erroneous
  - Time consuming to obtain accurate phone/articulatory features alignments.
- Phonetic transcriptions omit:
  - Different articulatory gestures
  - Co-articulation effects
  - Transition state between vowels and adjacent consonant
  - Anticipatory co-articulation
Continuous articulatory features

- Features that represent measured or simulated movement of vocal tract articulators.
  - X-ray microbeam
  - Electromagnetic Midsaggital Articulography (EMA)
  - Real Time Magnetic Resonance Imaging (rt-MRI)

Tract Variables

- Tract variable trajectories (TVs) are the temporal constriction dynamics
- TVs are relative measures of constriction produced by the moving articulators

<table>
<thead>
<tr>
<th>Constriction organ</th>
<th>Vocal tract variables</th>
<th>Articulators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lip</td>
<td>Lip Aperture (LA)</td>
<td>Upper lip, lower lip, jaw</td>
</tr>
<tr>
<td></td>
<td>Lip Protrusion (LP)</td>
<td></td>
</tr>
<tr>
<td>Tongue Tip</td>
<td>Tongue tip constriction degree (TTCD)</td>
<td>Tongue body, tip, jaw</td>
</tr>
<tr>
<td></td>
<td>Tongue tip constriction location (TTCL)</td>
<td></td>
</tr>
<tr>
<td>Tongue Body</td>
<td>Tongue body constriction degree (TBCD)</td>
<td>Tongue body, jaw</td>
</tr>
<tr>
<td></td>
<td>Tongue body constriction location (TBCL)</td>
<td></td>
</tr>
<tr>
<td>Velum</td>
<td>Velum (VEL)</td>
<td>Velum</td>
</tr>
<tr>
<td>Glottis</td>
<td>Glottis (GLO)</td>
<td>Glottis</td>
</tr>
</tbody>
</table>
Converting real articulatory data to TVs

- **Why?**
  - TVs are a more speaker independent representation than pellet positions.
    - Because they are relative measures of constriction position and degree
  - Use the TADA theoretical framework to analyze the phonological phenomena.
  - Less non-uniqueness in acoustic to articulatory mapping

### Methods to convert pellet trajectories from XRMB database to TVs (Nam H., et.al. 2012)

- **TADA (articular articulatory):**
  - Relative to ATR (articulatory articulation)
  - Tract variables are independent of speaker dimensions
  - Less non-uniqueness in acoustic to articulatory mapping

### Pellet positions for X-ray microbeam data

<table>
<thead>
<tr>
<th>Pellet</th>
<th>X-ray Microbeam data</th>
</tr>
</thead>
<tbody>
<tr>
<td>UL</td>
<td>Upper lip</td>
</tr>
<tr>
<td>LL</td>
<td>Lower lip</td>
</tr>
<tr>
<td>T1</td>
<td>Tongue tip</td>
</tr>
<tr>
<td>TR</td>
<td>Tongue body</td>
</tr>
<tr>
<td>T4</td>
<td>Tongue dorsum</td>
</tr>
<tr>
<td>MANI</td>
<td>Jaw</td>
</tr>
</tbody>
</table>

### Relative Tract Variable measures

<table>
<thead>
<tr>
<th>Tract Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA</td>
<td>Lip Aperature</td>
</tr>
<tr>
<td>LP</td>
<td>Lip projection</td>
</tr>
<tr>
<td>TBCL</td>
<td>Tongue body constriction location</td>
</tr>
<tr>
<td>TBCD</td>
<td>Tongue body constriction degree</td>
</tr>
<tr>
<td>TICL</td>
<td>Tongue Tip constriction location</td>
</tr>
<tr>
<td>TTCD</td>
<td>Tongue Tip constriction degree</td>
</tr>
</tbody>
</table>

### TADA theoretical framework to analyze the phonological phenomena.
Method to convert EMA sensor positions to TVs (Ghosh P.K. et.al. 2011)

- Assuming palate trace is not available
- These transformations were proposed for the MOCHA dataset
- \( LA = |UL_y - LL_y| \)
- \( PRO = |UL_x - LL_x| \)
- \( JAW\text{open} = |bn_y - LI_y| \)
  - (\(bn\) is a ref sensor)
- \( TTCD = \frac{|a_{tt}TT_x + b_{tt}TT_y + c_{tt}|}{\sqrt{a_{tt}^2 + b_{tt}^2}} \)
  - \(a_{tt}x + b_{tt}y + c_{tt} = 0\) represents the palate near the TT region
- TBCD can be similarly defined.

Method to convert EMA sensor positions to TVs (Sivaraman G. et.al. 2015)

- **Lip Aperture**
  - \( LA = \sqrt{(LL_x - UL_x)^2 + (LL_z - UL_z)^2} \)
- **Lip Protrusion**
  - \( LP = LL_x - \text{median}_{\text{all utterances}}(LL_x(m)) \)
- **Jaw aperture**
  - \( JA = \sqrt{(LI_x - UL_x)^2 + (LI_z - UL_z)^2} \)
- **Tongue Tip Constriction Degree (TTCD)**
  - \( TTCD = \text{Min}\{\text{Dist}(TT, Pal)\} \)
- **Tongue Tip Constriction Location (TTCL)**
  - \( TTCL = -\text{median}_{\text{all utterances}}(TT_x(m)) \)}
None of the natural speech databases contain gestural info.

- Initial exploration using Synthetic Speech

Word orthography (e.g., ‘garlic’) or ARPABET (e.g., [g a a r l i x k])

100000 words taken from CMU dictionary

TADA/ HLSyn

1. Gestures
2. TVs

Synthetic Acoustic Speech

Acoustic to Articulatory Speech Inversion

- Challenges
  - Highly nonlinear and somewhat non-unique mapping from acoustics to articulations
  - Different possible representations of articulatory features for similar acoustic content

- Approaches
  - Gaussian Mixture Models (GMM) and Hidden Markov Models (HMMs) (Hiroya et al., 2004)
  - Mixture Density Networks (Richmond K., 2007)
  - Artificial Neural Networks (ANN) (Mitra, 2010)

```
phoneme
• • • p1 p2 p3 • • •

/articulatory HMM/

articulatory parameter vector

x : articulatory parameter vector

articulatory-to-acoustic mapping

y1 y2 y3 • • • acoustic parameter vector

\[ y_t = f_i(x_t) = A_i x_t + b_i \]

... If \( x_t \) is generated by the HMM state \( i \)
```

Generalized smoothness criterion for speech inversion (Ghosh P.K., Narayanan S. 2010)

- Codebook based approach
- Optimization cost function:
  \[
  J = \text{Smoothness cost} + \text{Prob(codebook vector | estimated position)} \times \text{Distance(codebook vector, estimated position)}
  \]
- **Smoothness cost = High-pass power of the estimated trajectory**
- **Distance metric = Euclidean distance between estimated position and codebook vector**
- **\( \text{Prob(codebook vector | estimated position)} = \frac{\text{Distance}^{-1}(\text{codebook vector, estimated position})}{\sum_i \text{Distance}^{-1}(\text{estimated position, codebook vector}_i)} \)**

In GSC, \( j \)-th (\( j=1,\ldots,D \)) articulator trajectory is estimated as follows (\( D: \text{Total number of articulatory features} \)):

\[
\left\{ x_n^j ; 1 \leq n \leq N \right\} = \arg \min_{\hat{\theta}^j} \left\{ \sum_n \left( x_n^j \cdot h^j[n] \right)^2 + C_j \sum_n \sum_l \left( x_n^j - \eta_n^{l,j} \right)^2 p_n^l \right\}
\]

**Articulator specific smoothness**  **Data proximity term**
**Generalized smoothness criterion for speech inversion**

In GSC, $j$-th ($j=1,\ldots,D$) articulator trajectory is estimated as follows (D: Total number of articulatory features):

$$\{x_n^j, 1 \leq n \leq N\} = \arg \min_{\{h_n^j\}} \left\{ \sum_n (x_n^j \cdot h_n^j)^2 + C_j \sum_n \sum_l \left[ x_n^j - \eta_n^{j,l} \right]^2 \right\}$$

$j$-th articulator specific high-pass filter

$l$-th probable value of the $j$-th articulatory feature at the $n$-th frame ($1 \leq l \leq L$)

$GSC$ can be used for both subject-dependent and subject-independent inversion

**Trajectory Mixture density networks (t-MDN) for speech inversion (Richmond K., 2007)**

*Figure adapted from Zen H., Senior A. 2014*
MLPG algorithm
- We aim to maximize $P(O|Q)$ w.r.t. $O$ where $Q$ is the sequence of Gaussians output by the MDN
- Trajectory $O = [o_1, o_2, \ldots, o_T]^T$ where $o_t = [c_t, \Delta c_t, \Delta \Delta c_t]$ where $c_t$ is the static feature vector. In this case, $c = \text{Articulatory positions}$.
  - $O = WC$
- $W = \text{weights used to compute the deltas and double deltas}$
  $$\frac{\partial P(WC|Q)}{\partial C} = 0$$
  $$W^T U^{-1} WC = W^T U^{-1} M^T$$
- An expectation maximization (EM) algorithm is used to solve for $C$ in the above equation. (Tokuda et.al. 2000)

Distal Supervised Learning (Mitra et.al. 2010)
Deep Neural Network based Speech Inversion System (Mitra et.al. 2014; Sivaraman G. 2017)

- Multi-dimensional regression problem
- Artificial neural networks (ANN) suitable for the highly non-linear and non-unique mapping from acoustics to TVs (Mitra V. et al. 2010)
- Adam optimization technique used with stochastic gradient descent for training the DNNs

Results of Speaker Independent Speech Inversion

- **X-ray Microbeam (XRMB) dataset** converted to TVs
  - 46 different speaker (21 males, 25 females)
  - Around 4 hours of training data
- Trained on 36 speakers. Cross validation and testing on 5 speakers’ data for each
- Mean, Variance Normalization: Speaker specific normalization (SPKNORM)
- Pearson Product Moment Correlation (PPMC) between estimated and actual TVs

<table>
<thead>
<tr>
<th>TV</th>
<th>PPMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA</td>
<td>0.799</td>
</tr>
<tr>
<td>LP</td>
<td>0.670</td>
</tr>
<tr>
<td>TBCL</td>
<td>0.874</td>
</tr>
<tr>
<td>TBCD</td>
<td>0.749</td>
</tr>
<tr>
<td>TTCL</td>
<td>0.765</td>
</tr>
<tr>
<td>TTCD</td>
<td>0.864</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.787</strong></td>
</tr>
</tbody>
</table>
Example test utterance

“Combine all the ingredients in a large bowl”

Lip Aperture (LA)

Tongue Body Constriction Degree (TBCD)

Tongue Tip Constriction Degree (TTCD)

Lenited /l/

Bidirectional recurrent neural networks for speech inversion (Zhu P. et al. 2015)

- Experiments performed on single speaker MNGU0 dataset
- Acoustic-to-articulatory mapping

<table>
<thead>
<tr>
<th>Feature/Node</th>
<th>64</th>
<th>128</th>
<th>256</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSF</td>
<td>0.984</td>
<td>0.889</td>
<td>0.901</td>
</tr>
<tr>
<td>MFCC</td>
<td>0.599</td>
<td>0.565</td>
<td>0.585</td>
</tr>
</tbody>
</table>

- Text-to-articulatory mapping

<table>
<thead>
<tr>
<th>Features</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic + POS&amp;TOBI</td>
<td>1.870</td>
</tr>
<tr>
<td>Basic</td>
<td>1.925</td>
</tr>
<tr>
<td>Word2vec</td>
<td>2.530</td>
</tr>
<tr>
<td>Triphone2vec</td>
<td>2.348</td>
</tr>
<tr>
<td>Basic + Word2vec</td>
<td>1.884</td>
</tr>
<tr>
<td>Basic + Triphone2vec</td>
<td>1.881</td>
</tr>
<tr>
<td>Basic + POS &amp; TOBI + Word2vec</td>
<td>1.782</td>
</tr>
<tr>
<td>Basic + POS &amp; TOBI + Triphone2vec</td>
<td>1.734</td>
</tr>
</tbody>
</table>

Articulatory trajectories

BLSTM-2

BLSTM-1

Feedforward layer

Acoustic features

MFCC or LSF

OR

Text features

Word2vec, Triphone2vec, POS …
Articulatory gestures

- The TADA model is based on the theory of Articulatory Phonology (Browman and Goldstein, 1992) that defines speech as a constellation of coordinated articulatory gestures.
- Articulatory gestures are constricting actions for distinct organs/constrictors (lips, tongue tip, tongue body, velum and glottis) along the vocal tract.
- Each gesture is dynamically coordinated with a set of appropriate articulators. A word can be defined as a constellation of distinct gestures (gestural scores).
- Given the ARPABET transcription of an English word, the TADA model computes the gestural scores along with the inter-articulatory gestural coordination to produce the word and outputs the time functions of the vocal tract variables (TVs: degree and location variables of the constrictors) and model articulator variables.
Estimating articulatory gestures from speech (Mitra et.al. 2012)

Approach-1

- Cascaded ANN: Autoregressive ANN for gestural activation detection & Feed-forward ANN for gestural parameter detection
- Separate models for each gesture
- Optimal context windows: 90-190ms for gesture activation detection & 210-290ms for gestural parameter estimation

<table>
<thead>
<tr>
<th>Gesture Recognition accuracy (frame-wise)</th>
<th>93.66%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>89.62%</td>
</tr>
<tr>
<td></td>
<td>90.20%</td>
</tr>
</tbody>
</table>

Example estimate of gestures from speech+TV
Effects of Speaking rate on speech inversion

EMA-IEEE dataset (Tiede et.al. 2017)
- EMA data was collected from 8 subjects producing 720 Harvard sentences at 2 speaking rates.

- Speaking rate of Fast utterances ~ 0.66 times normal utterances

- Recorded EMA sensor positions were converted to TVs
  - We have 3 new additional TVs in the EMA-IEEE dataset (Jaw Aperture [JA], Tongue Root Constriction Degree [TRCD], and Location [TRCL] due to the presence of additional sensors in the EMA data

- Feedforward neural network based speaker dependent (SD) and speaker independent (LOO) speech inversion systems trained for each speaker and speaking rate.

- Same utterances used for training, crossval and testing for all systems to make a fair comparison

Results of speech inversion at different speaking rates

- Several Speaker dependent, rate dependent speech inversion systems were trained

<table>
<thead>
<tr>
<th>System name</th>
<th>Description</th>
<th>Amount of training data</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD_N</td>
<td>Speaker dependent Normal rate</td>
<td>~ 25 min</td>
</tr>
<tr>
<td>SD_F</td>
<td>Speaker dependent Fast rate</td>
<td>~ 18 min</td>
</tr>
<tr>
<td>SD_all</td>
<td>Speaker dependent all utterances</td>
<td>~ 33 min</td>
</tr>
<tr>
<td>LOO_N</td>
<td>Leave one speaker out Normal rate</td>
<td>~ 2.9 hours</td>
</tr>
<tr>
<td>LOO_F</td>
<td>Leave one speaker out Fast rate</td>
<td>~ 2.0 hours</td>
</tr>
<tr>
<td>LOO_all</td>
<td>Leave one speaker out all utterances</td>
<td>~ 4.9 hours</td>
</tr>
</tbody>
</table>

- Matched speaking rate results for different Speaker dependent, speaker independent; speaking rate specific, and mixed rate systems
Cross speaking rate evaluations

- Systems trained on normal rate were evaluated on fast utterances and vice versa
- Speaker dependent and independent tests were performed

<table>
<thead>
<tr>
<th></th>
<th>Normal rate test set</th>
<th>Fast rate test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD_N</td>
<td>0.823</td>
<td>0.749</td>
</tr>
<tr>
<td>SD_F</td>
<td>0.756</td>
<td>0.757</td>
</tr>
<tr>
<td>LOO_N</td>
<td>0.716</td>
<td>0.621</td>
</tr>
<tr>
<td>LOO_F</td>
<td>0.672</td>
<td>0.645</td>
</tr>
</tbody>
</table>

- Model trained on normal rate performs almost equivalent (avg. correlation lower by 0.008) to fast rate trained models on Fast utterances.
- Model trained on fast rate performs poorly on normal utterances (avg. correlation lower by 0.067)

Why bother about accuracy of the estimated TVs?

- Stop consonant place of articulation estimation on the XRMB dataset.
- Segment level classification of stop consonants into bilabial, alveolar, and velar.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groundtruth_TV</td>
<td>93.84%</td>
</tr>
<tr>
<td>MFCC</td>
<td>91.04%</td>
</tr>
<tr>
<td>XRMB_TV</td>
<td>84.07%</td>
</tr>
<tr>
<td>MFCC+XRMB_TV</td>
<td>92.54%</td>
</tr>
</tbody>
</table>

- In this task Groundtruth TV outperform MFCC features.
- Higher the TV estimation accuracy, better will be the classification performance
- Real TV combined with acoustic features outperform the combination of estimated TV and acoustics for phone classification. (Frankel J. and King S., 2001) (Ghosh P.K. and Narayanan S., 2010)
Speaker dependent inversion systems

- 10 speakers (5 males, 5 females) chosen randomly from the XRMB dataset such that each speaker had roughly same amount of data
- ANN Speech inversion systems trained for each speaker.
- Performance evaluated using PPMC measure on test set (same speaker) (10% of available data)

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Spk ID</th>
<th>Gender</th>
<th>Minutes of data</th>
<th>LA</th>
<th>LP</th>
<th>TBCL</th>
<th>TBCD</th>
<th>TTCL</th>
<th>TTCD</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker 1</td>
<td>JW12</td>
<td>M</td>
<td>7.01</td>
<td>0.78</td>
<td>0.75</td>
<td>0.89</td>
<td>0.79</td>
<td>0.73</td>
<td>0.87</td>
<td>0.80</td>
</tr>
<tr>
<td>Speaker 2</td>
<td>JW14</td>
<td>F</td>
<td>6.26</td>
<td>0.76</td>
<td>0.60</td>
<td>0.90</td>
<td>0.75</td>
<td>0.80</td>
<td>0.85</td>
<td>0.78</td>
</tr>
<tr>
<td>Speaker 3</td>
<td>JW24</td>
<td>M</td>
<td>7.58</td>
<td>0.80</td>
<td>0.70</td>
<td>0.88</td>
<td>0.69</td>
<td>0.72</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>Speaker 4</td>
<td>JW26</td>
<td>F</td>
<td>8.54</td>
<td>0.75</td>
<td>0.78</td>
<td>0.89</td>
<td>0.74</td>
<td>0.77</td>
<td>0.87</td>
<td>0.80</td>
</tr>
<tr>
<td>Speaker 5</td>
<td>JW27</td>
<td>F</td>
<td>6.83</td>
<td>0.74</td>
<td>0.69</td>
<td>0.85</td>
<td>0.69</td>
<td>0.66</td>
<td>0.82</td>
<td>0.74</td>
</tr>
<tr>
<td>Speaker 6</td>
<td>JW31</td>
<td>F</td>
<td>7.99</td>
<td>0.82</td>
<td>0.73</td>
<td>0.91</td>
<td>0.83</td>
<td>0.72</td>
<td>0.89</td>
<td>0.82</td>
</tr>
<tr>
<td>Speaker 7</td>
<td>JW40</td>
<td>M</td>
<td>6.32</td>
<td>0.76</td>
<td>0.53</td>
<td>0.91</td>
<td>0.72</td>
<td>0.82</td>
<td>0.86</td>
<td>0.77</td>
</tr>
<tr>
<td>Speaker 8</td>
<td>JW45</td>
<td>M</td>
<td>7.35</td>
<td>0.79</td>
<td>0.69</td>
<td>0.87</td>
<td>0.75</td>
<td>0.77</td>
<td>0.81</td>
<td>0.78</td>
</tr>
<tr>
<td>Speaker 9</td>
<td>JW54</td>
<td>F</td>
<td>6.76</td>
<td>0.69</td>
<td>0.51</td>
<td>0.87</td>
<td>0.71</td>
<td>0.84</td>
<td>0.80</td>
<td>0.74</td>
</tr>
<tr>
<td>Speaker 10</td>
<td>JW59</td>
<td>M</td>
<td>6.57</td>
<td>0.75</td>
<td>0.72</td>
<td>0.90</td>
<td>0.78</td>
<td>0.78</td>
<td>0.85</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Cross Speaker performance (Sivaraman G. 2017)

- Each speaker dependent system was evaluated on the test sets of the remaining 9 speakers.

<table>
<thead>
<tr>
<th>Spkr 1 (M)</th>
<th>Spkr 2 (F)</th>
<th>Spkr 3 (M)</th>
<th>Spkr 4 (F)</th>
<th>Spkr 5 (F)</th>
<th>Spkr 6 (F)</th>
<th>Spkr 7 (M)</th>
<th>Spkr 8 (F)</th>
<th>Spkr 9 (F)</th>
<th>Spkr 10 (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
<td>0.52</td>
<td>0.62</td>
<td>0.47</td>
<td>0.36</td>
<td>0.30</td>
<td>0.52</td>
<td>0.58</td>
<td>0.52</td>
<td>0.56</td>
</tr>
<tr>
<td>0.53</td>
<td>0.78</td>
<td>0.50</td>
<td>0.61</td>
<td>0.56</td>
<td>0.49</td>
<td>0.33</td>
<td>0.41</td>
<td>0.60</td>
<td>0.47</td>
</tr>
<tr>
<td>0.62</td>
<td>0.51</td>
<td>0.75</td>
<td>0.44</td>
<td>0.29</td>
<td>0.24</td>
<td>0.53</td>
<td>0.58</td>
<td>0.50</td>
<td>0.61</td>
</tr>
<tr>
<td>0.55</td>
<td>0.59</td>
<td>0.47</td>
<td>0.80</td>
<td>0.56</td>
<td>0.56</td>
<td>0.31</td>
<td>0.39</td>
<td>0.57</td>
<td>0.41</td>
</tr>
<tr>
<td>0.44</td>
<td>0.55</td>
<td>0.29</td>
<td>0.59</td>
<td>0.74</td>
<td>0.64</td>
<td>0.13</td>
<td>0.17</td>
<td>0.50</td>
<td>0.34</td>
</tr>
<tr>
<td>0.30</td>
<td>0.44</td>
<td>0.17</td>
<td>0.51</td>
<td>0.54</td>
<td>0.82</td>
<td>0.17</td>
<td>0.15</td>
<td>0.35</td>
<td>0.18</td>
</tr>
<tr>
<td>0.53</td>
<td>0.38</td>
<td>0.30</td>
<td>0.21</td>
<td>0.20</td>
<td>0.77</td>
<td>0.56</td>
<td>0.41</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>0.58</td>
<td>0.49</td>
<td>0.60</td>
<td>0.38</td>
<td>0.23</td>
<td>0.13</td>
<td>0.59</td>
<td>0.78</td>
<td>0.52</td>
<td>0.60</td>
</tr>
<tr>
<td>0.55</td>
<td>0.59</td>
<td>0.49</td>
<td>0.61</td>
<td>0.51</td>
<td>0.42</td>
<td>0.42</td>
<td>0.46</td>
<td>0.74</td>
<td>0.49</td>
</tr>
<tr>
<td>0.56</td>
<td>0.46</td>
<td>0.54</td>
<td>0.40</td>
<td>0.29</td>
<td>0.22</td>
<td>0.47</td>
<td>0.54</td>
<td>0.45</td>
<td>0.80</td>
</tr>
</tbody>
</table>
Speaker independent inversion using speaker independent articulatory HMMs (Hiroya S., Mochida T. 2006)

Remembering HMM based acoustic to articulatory inversion system …

phoneme

p1

p2

p3

…

articulatory HMM

\[ \lambda \text{ - articulatory parameter vector} \]

\( \begin{bmatrix} x_m \cdot & \sigma_m \end{bmatrix} \)

articulatory-to-acoustic mapping

\[ y = f(x) \]

\[ y_1 \]

\[ y_2 \]

\[ y_3 \]

…

acoustic parameter vector

Modeling inter-speaker variability in the articulatory HMM parameter domain

\[ \lambda^{AB} = \{ A^{AB} \cdot x + b^{AB}, H^{AB} \cdot \sigma_m \} \]

\[ \lambda^{BC} = \{ A^{BC} \cdot x + b^{BC}, H^{BC} \cdot \sigma_m \} \]

\[ \lambda^{AC} = \{ A^{AC} \cdot x + b^{AC}, H^{AC} \cdot \sigma_m \} \]

Maximum Likelihood Linear Regression (MLLR) based Speaker adaptation of the articulatory HMM parameters

Initial articulatory HMMs \( \lambda_{ini} \)

Speaker-adaptive matrix \( A^{(i)}, b^{(i)}, H^{(i)} \)

Speaker-independent articulatory HMMs \( \lambda_{Hi} \)

Generalized acoustic space and smoothness criterion based speaker-independent inversion (Ghosh P.K., Narayanan S. 2011)

\[ \{ x_n^i : 1 \leq n \leq N \} = \arg \min_{\{ x_n^i \}} \left[ \sum_n x_n^i \cdot h_j[n]^2 + C_j \sum_{t=1}^T \left( x_n^i - \eta_n^i \right) \cdot \left( p_n^i \right) \right] \]

Training

\( \{ x_n^i \} \)

Test

\( \{ u_n \} \)

Distance?
Vocal Tract Length Normalization based speaker Adaptation for speech inversion (Sivaraman G. et.al. 2016)

- **Objective:** To adapt the acoustic space to match the target speaker
- **Vocal Tract Length Normalization (VTLN)**
  - Method popularly used for acoustic model adaptation in ASR
  - Frequency domain warping to account for different tract lengths
  - Warping factor needs to be chosen to maximize acoustic similarity

Speaker acoustic spaces (Sivaraman G. et.al. 2016)

- 64 Gaussian mixtures to model each speaker’s acoustic space
- Gaussian Mixture Model (GMM) acoustic spaces are trained for each speaker in the dataset
Vocal Tract-length normalization (VTLN)

- Frequency domain warping like VTLN to adapt acoustic features of one speaker to another.
- Speaker $S_i$’s acoustic features are transformed to speaker $S_j$’s acoustic space using the VTLN parameter $\alpha_{ij}$ decided by a Maximum Likelihood criterion.

Speaker Adaptation scheme

- Unsupervised GMM Acoustic models are trained for each speaker (64 Gaussians per mixture model).
- Let the GMM acoustic model for speaker $S_j$ be $\lambda_j$
- The warped acoustic features for a frame of speech signal from speaker $S_i$ to the target speaker $S_j$ be $x_{ijt}$
- Most likely warping factor $\alpha_{ij} (S_i \rightarrow S_j)$ is given by –
  \[ \alpha_{ij} = \arg\max_{\alpha} \sum_{t=1}^{N} \log(P(x_{ijt}|\lambda_j)) \]
  - Best warping factor $\alpha_{ij}$ is obtained by grid search by sweeping the value from 0.8 to 1.2 in steps of 0.025
Speaker transformation of training dataset
(Sivaraman G. et al. 2016)

![Illustration showing VTLN adaptation for different speakers]

Improvement due to Vocal Tract Length Normalization (VTLN)

- Each test speaker's acoustic features are adapted to the target feature by a warping of the frequency spectrum motivated by VTLN.

Average corr = 0.55  
Average corr = 0.62

<table>
<thead>
<tr>
<th>Spk 1</th>
<th>Spk 2</th>
<th>Spk 3</th>
<th>Spk 4</th>
<th>Spk 5</th>
<th>Spk 6</th>
<th>Spk 7</th>
<th>Spk 8</th>
<th>Spk 9</th>
<th>Spk 10</th>
<th>Average corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td>0.84</td>
<td>0.85</td>
<td>0.86</td>
<td>0.87</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
<td>0.91</td>
<td>0.92</td>
<td>0.55</td>
</tr>
<tr>
<td>0.84</td>
<td>0.83</td>
<td>0.85</td>
<td>0.86</td>
<td>0.87</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
<td>0.91</td>
<td>0.92</td>
<td>0.62</td>
</tr>
<tr>
<td>0.85</td>
<td>0.84</td>
<td>0.85</td>
<td>0.86</td>
<td>0.87</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
<td>0.91</td>
<td>0.92</td>
<td>0.55</td>
</tr>
<tr>
<td>0.86</td>
<td>0.85</td>
<td>0.86</td>
<td>0.87</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
<td>0.91</td>
<td>0.92</td>
<td>0.93</td>
<td>0.62</td>
</tr>
<tr>
<td>0.87</td>
<td>0.86</td>
<td>0.87</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
<td>0.91</td>
<td>0.92</td>
<td>0.93</td>
<td>0.94</td>
<td>0.62</td>
</tr>
<tr>
<td>0.88</td>
<td>0.87</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
<td>0.91</td>
<td>0.92</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Greyscale values represent the test set correlation between estimated and actual TVs. White = 1, Black = 0.

Sivaraman G. et al., 2016)
Estimated TVs with and without speaker adaptation

Future directions for acoustic-to-articulatory inversion

- Effectively combining synthetic articulatory data with real articulatory measurements.
- Extend speaker normalization to include normalization in the articulatory domain
- Expand the resolution on EMA articulatory data using co-registered MRI datasets.
- Consolidating articulatory data collected using different methods for training a single speech inversion system.
- Combine discrete articulatory features with continuous articulatory data to improve estimation of both types of AFs.
Lexical units are phonemes and acoustic units are AFs

- Acoustic units are based on AFs or both AFs and phonemes (Metze and Waibel, 2002; Stüker et al., 2003; Juneja and Espy-Wilson, 2004; Livescu et al., 2007).
- Each acoustic unit is modeled with a GMM or with discriminative classifiers like ANNs or support vector machines.
- Each phoneme-based lexical unit is deterministically mapped to its AF attributes.
- The scores from different AF-based acoustic models are combined to arrive at the local emission score of HMM states.
- On continuous speech recognition and cross-lingual adaptation tasks, the use of AF-based acoustic models in combination with phoneme-based acoustic models resulted in a relative reduction in word error rate (WER) of about 5–10% compared to the use of phoneme-based acoustic models alone (Metze and Waibel, 2002; Stüker et al., 2003).

Lexical units are AFs and acoustic units are AFs

- Analogous to standard HMM-based ASR systems where both lexical and acoustic units are either based on context-independent or context-dependent subword units.
- However, the subword units are AFs determined from the AF-based pronunciation lexicon (Deng et al., 1997; Richardson et al., 2003; Kirchhoff, 1996; Wester et al., 2004; Livescu et al., 2007).
- The AF-based pronunciation lexicon transcribes each word in terms of the positions of the articulators.
- Each AF is associated with its own hidden state variable.
Approaches to Applying Articulatory features to ASR
(Rasipuram R., Magimai-Doss M., 2015)

- **Lexical units are phonemes and acoustic units are phonemes**
  - Similar to standard HMM-based ASR systems, both lexical and acoustic units are based on phonemes.
  - AF representations are used as auxiliary information to enhance the performance of the acoustic model (Kirchhoff et al., 2002; Siniscalchi et al., 2012, Mitra et al. 2011, Mitra et al., 2017).
  - Acoustic model can be viewed as a two-stage classifier.
    - Stage 1: Acoustic features → AFs
    - Stage 2: AFs → Phonemes OR AFs + Acoustic features → Phonemes
  - These systems have achieved a relative reduction in WER of about 5–6% on noise robust ASR tasks and cross-lingual ASR tasks compared to the systems where acoustic-to-phoneme information is directly modeled (Kirchhoff et al., 2002; Siniscalchi et al., 2012).

References

References

Robust Speech Recognition

Goal: To ensure a reasonable recognition accuracy despite mismatch in speech used for training & testing

Mismatch occurs in
- Acoustic characteristics
- Articulatory characteristics
- Phonetic characteristics

Mismatches are due to
- Coarticulation and lenition (rate, style, emotion)
- Noise contamination
- Speaker variation (gender, age, dialect, accent, etc.)
Speech variability: Coarticulation

Coarticulation: Influence of one phone upon another during speech production

- Coarticulation typically modeled by contextualized sub-word units
- Implicitly modeled through LSTM acoustic models in a data driven way
  - No formal approach used in practice to model coarticulation

Articulatory Gestures help to model coarticulation and lenition in a systematic way

- Gesture is a constriction action along the vocal tract (Browman & Goldstein, 1989)
- Speech can be decomposed into a constellation of Gestures
- Tract variable trajectories (TVs) are the temporal constriction dynamics

<table>
<thead>
<tr>
<th>Constriction organs</th>
<th>Vocal tract variables</th>
<th>Articulators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lip</td>
<td>Lip Aperture (LA)</td>
<td>Upper lip, lower lip, jaw</td>
</tr>
<tr>
<td></td>
<td>Lip Protrusion (LP)</td>
<td></td>
</tr>
<tr>
<td>Tongue Tip</td>
<td>Tongue tip constriction degree (TTCD)</td>
<td>Tongue body, tip, jaw</td>
</tr>
<tr>
<td></td>
<td>Tongue tip constriction location (TTCL)</td>
<td></td>
</tr>
<tr>
<td>Tongue Body</td>
<td>Tongue body constriction degree (TBCD)</td>
<td>Tongue body, jaw</td>
</tr>
<tr>
<td></td>
<td>Tongue body constriction location (TBCL)</td>
<td></td>
</tr>
<tr>
<td>Velum</td>
<td>Velum (VEL)</td>
<td>Velum</td>
</tr>
<tr>
<td>Glottis</td>
<td>Glottis (GLO)</td>
<td>Glottis</td>
</tr>
</tbody>
</table>

Tract variables (TVs)
Coarticulation and Gestures

**Articulatory Features (AFs) and Acoustic Modeling**

Variability in speech can be accounted for by incorporating speech production knowledge into ASR (Stevens, 1960)

- One of the earliest work to incorporate AFs in ASR was done by (Schmidbauer, 1989)
  - Used to recognize German speech using 19 AFs that described the manner and place of articulation.
  - AFs were used as input to a HMM acoustic model and an improvement of 4% was observed over the baseline.
  - AFs were found to be robust against speaker variability compared to the HMM-MFCC baseline.

- **Articulatory Features**: Heuristically defined, specifying manner & place of articulation (Kirchhoff, 1999)
  - Identified as pseudo-articulatory features.
  - Voiced/unvoiced, place and manner of articulation, lip-rounding, etc.
  - Demonstrated that AFs in combination with MFCCs provided increased recognition robustness against background noise.

Note: Due to rate of change of speech, the degree of overlap between the gestures is altered and there may be undershoot in the gestural target, but the overall gestural pattern remains the same.
Articulatory features and Acoustic Modeling

- **Articulatory trajectories**: Placing sensors (pellets) on articulators and tracking them (Frankel, 2000 & 2001)
  - Explored Linear Dynamic models and single hidden layer neural networks to learn acoustic to articulatory feature transforms.
  - Proposed a hybrid DBN/ANN acoustic model and reported gain in ASR performance for OGI number corpus.

- **Vocal tract resonances (VTR) and Vocal Tract Constrictions** (Deng et al., 1998, 2000, 2004)
  - Combined continuous and dynamic phonetic information of speech production with discrete phonological features.
  - Investigated statistical hidden dynamic model to account for phonetic reduction in conversational speech.
  - Demonstrated improved performance on Switchboard task compared to a GMM-HMM baseline.

Articulatory features and Acoustic Modeling

- **Dynamic Bayesian Networks** have been explored to incorporate temporal dependencies within and across articulators
  - (Livescu et al., 2003) reported significant reduction in WER on Aurora-2 corpus.
  - Similar improvements were observed in (Mitra et al., 2011), where both continuous articulatory features and discrete gestural units were used.

<table>
<thead>
<tr>
<th>WER from Aurora-2 speech recognition task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WER (%)</strong></td>
</tr>
<tr>
<td>HMM</td>
</tr>
<tr>
<td>MFCC</td>
</tr>
<tr>
<td>MFCC+TV+Gesture</td>
</tr>
<tr>
<td>G-DBN</td>
</tr>
<tr>
<td>DBN</td>
</tr>
<tr>
<td>MFCC</td>
</tr>
<tr>
<td>MFCC+TV</td>
</tr>
</tbody>
</table>

**Interspeech-2018**

Hyderabad, India

Page 86 of 103
Articulatory features and Acoustic Modeling

- Overlapping *Gesture*-like features (Sun & Deng, 2002)
  - Overlapping feature set was semi-manually transcribed for one speaker in X-ray microbeam database.
  - 1.8% absolute improvement in word recognition accuracy on TIMIT dataset.

- The usual trend has been to learn an articulatory transformation and use that in ASR.
  - Suffers from data scarcity
  - Expensive to obtain matched acoustic and articulatory pairs.
  - The inverse transform us difficult to learn
  - Amenable to domain and data mismatch.
  - To learn transforms for noisy and reverberation conditions, training data is augmented with simulated distortions (Mitra et al., 2017)

Articulatory features and Acoustic Modeling

- Deep Neural Networks have been explored for both speech inversion and acoustic model training (Badino et al., 2016)
  - Phone error rate reduction as high as 10% was reported for the MOCHA-TIMIT dataset

- Variational deep Canonical Correlation Analysis with Private variables (VCCAP) was proposed by (Tang et al., 2018)
  - Instead of deriving the articulatory features directly from speech VCCAP obtains an intermediate representation.
  - Such latent representations can generalize across domains and improve ASR performance.

- Articulatory representations provide ASR performance improvement when combined with traditional acoustic features.
  - How can we combine them well?
  - Simple solution: Feature combination
  - Acoustic features and articulatory representations have quite different properties.
Articulatory features and Acoustic Modeling

- Fused Convolutional – Deep neural Net (F-CNN/DNN) architecture has been found to be quite useful in combining (Mitra et. al., 2018)
  - Leverages the strengths of a filterbank energy based CNN architecture
  - Exploits the complementary properties of articulatory representations.

### LVCSR performance on SWB-2K CTS task

<table>
<thead>
<tr>
<th>Feature</th>
<th>Model</th>
<th>SWB WER</th>
<th>CH WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFB+fMLLR</td>
<td>CNN</td>
<td>10.8</td>
<td>19.2</td>
</tr>
<tr>
<td>DOC+fMLLR</td>
<td>CNN</td>
<td>10.4</td>
<td>18.1</td>
</tr>
<tr>
<td>MFB+fMLLR + DOC+fMLLR</td>
<td>CNN</td>
<td>9.8</td>
<td>17.2</td>
</tr>
<tr>
<td>MFB+fMLLR + DOC+fMLLR + TV</td>
<td>f-CNN-DNN</td>
<td>9.5</td>
<td>16.9</td>
</tr>
</tbody>
</table>

Articulatory trajectories for ASR

- ASR deals with an open set of speakers: speakers not seen during training
- Bottleneck for articulatory features:
  - Recorded articulatory data quite expensive to obtain.
  - Existing datasets contain limited number of speakers.
    - Calibration mismatch (sensor placement) can introduce anomaly in speaker variation.
    - Speakers may not be able to spontaneous speak and use articulators with transducers placed during recording time.
- Need: Speaker-invariant Articulatory representation for robust ASR acoustic modeling.

**Speech Inversion:** Model driven transform to estimate articulatory dynamics given speech signal as input.
What is Speech Inversion?

Speech Inversion suffers from non-linearity and non-uniqueness
- Non-linearity: Quantal nature of speech (equal changes in articulatory dynamics may not have equal changes in acoustics)
- Non-uniqueness: Different vocal tract configurations can yield similar acoustic realizations (e.g., bunched vs retroflex /r/)

Speech Inversion Latest trends
- Most initial speech inversion work focused on single speakers
  - Even with single speaker speech inversion has been a challenging problem given its ill-posed properties.
  - Non-uniqueness and Non-linearities.
- Given increase focus on the use of articulatory representations in speech based applications the focus has shifted to learning speaker-invariant transforms.
- **Multi-speaker speech inversion is challenging**
  - Promising results on real articulatory data reported in (Ghosh & Narayanan, 2011; Afshan & Ghosh, 2015; Sivaraman et al., 2016)
  - Deep Networks can learn multispeaker inversion using data generated by articulatory synthesizer (Mitra et. al., 2017)
Articulatory features, Acoustic Modeling and Deep Learning

Deep Learning jet-streamed speech inversion performance

- Fully connected multi-layered Neural Nets demonstrated impressive performance gains.
- Deeper Neural Nets demonstrated robustness and even better performance
  - GPU based training opened up the possibility of training models with 100’s of hours of data
  - More training data resulted in more resilient models.
- Convolutional Neural nets improved performance even more.
  - Robustness to noise and distortions in acoustic space.
  - Better performance
  - Simpler input acoustic features: filterbank features without Discrete Cosine Transforms
  - Relatively speaker invariant

Performance Gains with Neural Nets

![Speech Inversion performance with layer depth (Mitra et al., 2014)](image-url)
Performance Gains with Neural Nets

Speech Inversion performance: DNN vs CNN (clean data trained, tested on noisy data)

Speech Inversion performance: DNN vs CNN trained with noisy data (noisy data trained, tested on noisy data)
Speech Inversion: takeways from deep learning models

- Deeper models always gave better performance
  - Performance gain from layer depths tended to saturate after 3 to 5 hidden layers (for synthetically generated acoustic articulatory data pairs ~ 400 hrs)
- CNN always performed as-good-as if not better than DNNs.
- Data augmentation
  - Introducing speaker variation can significantly improve speech inversion robustness
  - Augmenting data with noise and reverberation corruption ruggedized the models for unseen acoustic conditions
- Use of robust acoustic features can help to boost speech inversion performance on unseen noisy conditions.

Speech Recognition Acoustic Models: Early DNNs

- DNN acoustic model was trained in (Mitra et al., 2014) taking articulatory trajectories as input.
- Compared MFCC and robust features both for speech inversion and acoustic model feature combination for Aurora-4 (8kHz) ASR task.
  - Contextualizing the articulatory features was found to be useful.
  - Provided robustness to noise and channel degradation.

<table>
<thead>
<tr>
<th></th>
<th>Matched Channel</th>
<th>Missmatched Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>33.1</td>
<td>39.0</td>
</tr>
<tr>
<td>MFCC+TV</td>
<td>29.9</td>
<td>35.5</td>
</tr>
<tr>
<td>NMCC</td>
<td>28.3</td>
<td>33.2</td>
</tr>
<tr>
<td>NMCC+TV</td>
<td>28.5</td>
<td>32.6</td>
</tr>
</tbody>
</table>

TV: Tract variable trajectories  
NMCC: Normalized Modulation cepstral coeff.
Speech Recognition Acoustic Models: Early DNNs

- DNN acoustic model was trained in (Mitra et al., 2014) taking articulatory trajectories as input.
- Compared MFCC and robust features both for speech inversion and acoustic model feature combination for Aurora-4 (8kHz) ASR task.
  - Contextualizing the articulatory features was found to be useful.
  - Provided robustness to reverberation.

<table>
<thead>
<tr>
<th></th>
<th>ROOM-1</th>
<th>ROOM-2</th>
<th>ROOM-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOCC</td>
<td>12.83</td>
<td>13.99</td>
<td>16.81</td>
</tr>
<tr>
<td>MFCC</td>
<td>8.64</td>
<td>11.14</td>
<td>14.08</td>
</tr>
<tr>
<td>DOCC+MFCC+TV</td>
<td>13.27</td>
<td>14.08</td>
<td>16.81</td>
</tr>
</tbody>
</table>

DOCC: Damped Oscillator Cepstral Coeff.

Traditional Ways of using Articulatory features

Typically articulatory trajectories have been combined with acoustic features
Assumption: The acoustic space and the articulatory space has similar properties

Caveat:
- Acoustic features (spectral energies) have spatial correlation
  - Great candidate for convolutional neural nets.
- Articulatory features are typically low pass in nature
  - May or may not correlate with other articulatory dimensions.
- Why not learn the two spaces separately?
Hybrid Convolutional Net: Results

Two separate convolutional nets (acoustic model and articulatory model) sharing the same output layer

Hybrid convolutional network (HCNN)

Hybrid Convolutional Net: Results

Two separate convolutional layers (acoustic-space convolution and articulatory-space convolution) sharing same hidden layers

Fused convolutional network (fCNN)
Hybrid Convolutional Net

Two separate convolutional nets (acoustic model and articulatory model) sharing the same output layer

<table>
<thead>
<tr>
<th></th>
<th>WER from Aurora-4 evaluation task</th>
<th>WER from WSJ evaluation task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CNN GFB</td>
<td>CNN GFB</td>
</tr>
<tr>
<td></td>
<td>9.4</td>
<td>5.7</td>
</tr>
<tr>
<td></td>
<td>HCNN GFB+TV-DNN</td>
<td>HCNN GFB+TV-CNN</td>
</tr>
<tr>
<td></td>
<td>9.2</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>HCNN GFB+TV-CNN</td>
<td>HCNN GFB+TV-CNN</td>
</tr>
<tr>
<td></td>
<td>9.1</td>
<td>5.6</td>
</tr>
</tbody>
</table>

**GFB**: Gammatone filter-bank energy

Hybrid Convolutional Net v/s Fused Convolutional Net

Performance comparison of fCNN and HCNN for Aurora-4 task

<table>
<thead>
<tr>
<th></th>
<th>WER from Aurora-4 ASR task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CNN GFB</td>
</tr>
<tr>
<td></td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>HCNN GFB+TV-DNN</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>HCNN GFB+TV-CNN</td>
</tr>
<tr>
<td></td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td>fCNN GFB+TV-DNN</td>
</tr>
<tr>
<td></td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td>fCNN GFB+TV-CNN</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
</tbody>
</table>

**GFB**: Gammatone filter-bank energy
Hybrid Convolutional Net v/s Fused Convolutional Net

Performance comparison of fCNN and HCNN for WSJ task

<table>
<thead>
<tr>
<th></th>
<th>WER from WSJ ASR task</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN MFB</td>
<td>6.7</td>
</tr>
<tr>
<td>DNN GFB</td>
<td>6.4</td>
</tr>
<tr>
<td>CNN GFB</td>
<td>5.7</td>
</tr>
<tr>
<td>HCNN GFB+TV-DNN</td>
<td>5.4</td>
</tr>
<tr>
<td>fCNN GFB+TV-DNN</td>
<td>5.6</td>
</tr>
</tbody>
</table>

**GFB:** Gammatone filter-bank energy

LVCSR: Switchboard CTS (2000 hrs)

Fused CNN-CNN architecture: Faster to train and fewer parameters

Fused CNN DNN (fCNN-DNN) architecture
LVCSR: Switchboard CTS (2000 hrs)

- Kaldi hybrid DNN-HMM ASR system.
- 5.6K context-dependent (CD) states
- Feature splicing: 17 frames
- Trained by using cross-entropy (CE)
  - followed by sequence training using maximum mutual information (MMI) criterion
- Acoustic features: 200 convolutional filters of size 8
- Articulatory features: feed-forward layer with 100 neurons.
- Pooling size of 3
- Fully connected network had five hidden layers, with 2048 nodes per hidden layer.
- Training: Initial four iterations with a constant learning rate of 0.008, followed by learning-rate halving based on the CV error.
- SGD with a mini-batch of 256.

LVCSR: Switchboard CTS (2000 hrs)

- WERs from NIST 2000 evaluation set.
Articulatory Features and Speech impairment

- Certain health conditions can result in speech impairment, such as Dysarthria
  - Dysarthria impacts speech articulation mechanism.
  - Produces slurred and often unintelligible speech.
  - Deviation of spectral characteristics in the acoustic signal results in ASR performance errors.
  - Speakers use the same speech production mechanism, however articulation may vary from normal speech

Questions:
- Can articulatory representations help in detecting a speaker’s effort in producing certain sounds?
- Can articulatory features compliment acoustic features under such adverse cases?

(Emre et al., 2018) explored the use of articulatory features for recognizing speech from dysarthric speakers.

Dysarthric Speech Recognition

Dysarthric data collected from Dutch and Flemish speakers (more details in (Emre et al., 2018))

<table>
<thead>
<tr>
<th></th>
<th>WER from Dutch Dysarthric test set</th>
<th>WER from Flemish Dysarthric test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFB DNN</td>
<td>22.9</td>
<td>36.4</td>
</tr>
<tr>
<td>MFB CNN</td>
<td>21.1</td>
<td>33.5</td>
</tr>
<tr>
<td>MFB+TV fCNN</td>
<td>19.1</td>
<td>32.2</td>
</tr>
</tbody>
</table>
Articulatory Features and Mental Health

- Mental Health conditions affect the speech production system
  - Depression can be life threatening and can degrade quality of life
  - Detection, evaluation and early treatment can cure depression
  - Current diagnosis is purely subjective evaluation based
  - Can often be expensive and time-consuming
  - Automatic detection of depression can facilitate quick evaluation and treatment.
- Speech can be used for automatic depression detection.
  - Mental health conditions result in deviation of speech articulation compared to canonical speech.
  - Articulatory Features have been investigated as a possible cue for predicting depression from speech (Mitra et al., 2015).

Articulatory Features and Mental Health

  - AVEC-2014: Audio-visual depression corpus of 84 speakers
  - Three equal subsets of train, dev and test.
    - Two splits: (a) read and (b) spontaneous speech.
    - Spontaneous speech more relevant for mental health detection (Mitra et al. 2015)
  - Self reported scores in Beck Depression rating scale
    - Integer numbers between 0 to 44.
- Performance measured using:
  - Mean absolute error (MAE)
  - Root mean squared error (RMSE) and
  - Pearson’s product moment correlation (rPPMC)
Articulatory Features for Depression Prediction

- Performance on AVEC-2014 dataset

<table>
<thead>
<tr>
<th>Feature</th>
<th>MAE</th>
<th>RMSE</th>
<th>r_PPMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>7.99</td>
<td>10.00</td>
<td>0.58</td>
</tr>
<tr>
<td>GCC</td>
<td>7.53</td>
<td>9.45</td>
<td>0.65</td>
</tr>
<tr>
<td>Pitch</td>
<td>8.13</td>
<td>10.53</td>
<td>0.49</td>
</tr>
<tr>
<td>Trajectory</td>
<td>7.11</td>
<td>9.33</td>
<td>0.63</td>
</tr>
<tr>
<td>AF</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Articulatory Features for Mental health detection

- AFs performed quite competitively w.r.t MFCC, GCC and Pitch features.
  - Provided lower errors (in terms of RMSE and MAE)
  - Competitive correlation w.r.t the ground truth scores.
  - AFs had 8 dimensions compared to 13 dimensional MFCC and GCC features.

TODO:
- Has the potential to perform even better if AF models are trained with real articulatory data.
- Modeling AF dynamics can capture slurry speech, incomplete articulation, lack of rhythmicity etc.
Future Directions

- Better Articulatory Models result in better performance
  - LSTM network can learn speech to articulation mapping while leveraging input output context.
  - Implicitly learn smoothness criteria of AF features
    - Articulatory movements are low-pass in nature (cite)
- Learning articulatory and acoustic spaces separately for ASR is found to be quite useful
  - Explore similar extension with CTC and Attention modeling techniques
  - Investigate multi task learning:
    - Learn speech inversion while training acoustic modeling for ASR.

Future Directions

- Open Question: Is it really necessary to train speech inversion models
  - Intermediate representations have been quite useful in acoustic model training.
  - Quite useful for transferring information across multiple domains.

Caveat: Lack of interpretability

- Use canonical Articulatory information where interpretability is useful
  - Example: mental health detection, speaker recognition etc.
- Use hidden representations for boosting model performance
  - Example: Cross-domain ASR, keyword detection etc.


References


