Measuring and Using Speech Production Information: Some new opportunities

Shrikanth (Shri) Narayanan
Signal Analysis and Interpretation Laboratory (SAIL)
http://sail.usc.edu

University of Southern California
ISCA: International Speech Communication Association

ISCA started in 1999 jointly with
ESCA (European Speech Communication Association):
ICSLP (International Conference of Spoken Language Processing)

• Purpose:
to promote Speech Communication Science and Technology, both in the industrial and academic areas, covering all the aspects of Speech Communication: acoustics, phonetics, phonology, linguistics, natural language processing, artificial intelligence, cognitive science, signal processing, pattern recognition, etc.

• ISCA offers a wide range of services; in particular Interspeech, ISCA workshops, SIGs (special interest groups), and Distinguished Lectures.
ISCA Objectives:

• to stimulate scientific research and education,
• to organize conferences, courses and workshops,
• to publish, and to promote publication of scientific works,
• to promote the exchange of scientific views in the field of speech communication,
• to encourage the study of different languages,
• to collaborate with all related associations,
• to investigate industrial applications of research results,
• and, more generally, to promote relations between public and private, and between science and technology.
The Speech Chain

Pinson & Denes: Adapted from Rabiner & Schafer, 2010,
What role can speech science play in speech technology development?
Lecture Premise & Layout

Understanding the system that produces speech is essential to advancing speech science and improving speech technology

- Scientific studies: Empirical analyses, Direct system (forward) modeling
- Technology studies: Feature engineering, Inverse modeling, Applications to ASR, Speaker Modeling, Synthesis, Clinical problems

- Measuring speech production
  - Multimodal approaches: EMA, Ultrasound, MRI,..

- Extracting features (representations)
  - Direct & Estimated (inversion)

- Modeling speech production
  - Theoretically inspired & Data-driven

- Applications
  - ASR, Speaker modeling
Sound out

Larynx vibrates

Constrictions are formed

Air in
Diverse Stimuli

- Vowels, Continuants
- Read sentences
- Spontaneous
- Non speech gestures

Multimodal Data Acquisition

- RT-MRI
- 3d MRI
- Audio
- EMA

Multimodal Analysis & Modeling

- direct image analysis
- forced alignment
- articulator tracking
- acoustic feature extraction
- cross-modal registration
- airway segmentation
- morphological characterization
- task-dynamic modeling
- dynamic 3d vocal tract modeling
- joint factor analysis, manifold learning, multiview learning

Scientific Insights, Models, Theory

- dynamics of production
- 3d vocal tract shaping
- articulatory coordination
- source-filter interaction
- realization of prosody
- speaker-specific phonetics
Speech Production Studies: Data Is Integral

- Observe, measure, visualize articulatory details during speech

- Long history of instrumentation and imaging applications

- Number of techniques, each with its own strengths and limitations
  - Spatial and temporal resolution
  - Subject safety
  - Flexibility, ease of use, portability
  - Data interpretability
  - Specific research and application needs
Commonly used speech production data types

**X-ray**
- high temporal and spatial resolution
- radiation; limited resolution for soft tissue

**Electromagnetometry (EMA)**
- safe; high temporal resolution; flesh point tracking
- invasive; spatially sparse data; not for pharyngeal structures

**Ultrasound**
- safe; high temporal resolution; portable
- provides incomplete view of vocal tract

**Palatography**
- safe; high temporal resolution; portable
- invasive; provides indirect information on oral cavity
Classic Speech Production Data Examples

X-ray (Stevens, 1962)
http://psyc.queensu.ca/~munhallk/05_database.htm

Ultrasound (Stone, 1980)
http://www.speech.umd.edu

Electropalatography
(courtesy: UCLA Phonetics Lab)
ELECTROMAGNETIC ARTICULOGRAPHY (EMA)

A. Wrench, A multichannel articulatory database and its application for automatic speech recognition
Proceedings 5th Seminar of Speech Production, 2000
Transducer placement and Audio

Subject/Interlocutor microphones
- experimenter/interlocutor: close talking and throat mic
- subject: shot gun mic

Audio recorded synchronous to articulatory data
Magnetic Resonance Imaging (MRI)

- Static scenario
- Dynamic scenario
  - Cine MRI
  - Tagged Cine MRI
  - Real time MRI
MRI for structural vocal tract imaging

Capable of 3D imaging of the hydrogen concentration in human body

Number of advantages:

– Non-invasive, no ionizing radiation
– Arbitrary scan plane: Information on complete vocal tract geometry
– Excellent, flexible structural differentiation: Good soft tissue contrast, SNR
– Amenable to computerized 3D modeling: reconstruction and visualization
– Quantitative information: area function and acoustic relations
– Variability analyses

Limitations/Challenges

– Slow: Spatial & Temporal resolution tradeoffs, optimizing to a given application
– Noisy images: Susceptibility, blurring artifacts
– Imaging teeth
– Interaction with other physiological activities: respiration, swallowing, other movement
– Clean, Synchronized audio (and other modalities, as needed)
– Ease of experimentation, including cost and portability
MRI for structural vocal tract imaging

Progressive improvements in overcoming these challenges
- Temporal and spatial resolution through various pulse sequence
- Gradient echos, Turbo Spin Echo, Spoiled GRASS; receiver coils
- Ingenious ways to capture teeth
- New audio acquisition strategies
- Advanced signal and image processing & statistical modeling methods

Applications to support both new
- Scientific inquiry and
- Technology development
MRI: Static Vocal tract Information

Information on 3D vocal tract structure/shape, inter-speaker variations
Accurate vocal tract measurements: area functions, length
Detailed studies on vowels and a number of continuant speech sounds
Facilitated new acoustic modeling studies
  – Vowels, Nasals, Fricatives, Liquids: English and other languages

Midsagittal vowel images from Haskins (from Goldstein)

3D airway reconstruction for vowel /a/, Univ. Iowa (Story)
http://everest.radiology.uiowa.edu/nlm/app/vocal/vocal.html
Consonant Example: Fricatives in English

3D Vocal tract and tongue shapes for /sh/

MRI-derived area functions for fricatives

Modeling fricative acoustic spectra

Strident fricative spectra derived from hybrid source model inputs using the parametric dipole spectra: dashed (model) solid (natural speech)

Dynamic MRI

BUT............
Dynamic MRI

• Speech is an inherently dynamic process

• Term “Dynamic” in speech imaging context used to refer to
  – Acquisition of images from actively articulating subject rather than static postures
  – The imaged object property: moving speech articulators
  – Different ways of capturing speech motion: trading off spatial and temporal resolution cleverly

• Reconstructions from repetitions ("cine" MRI)
• Increased temporal sampling ("real time" MRI)
MRI acquisition: spatial frequency domain

E.g. 10 ms/line => 128 x 10 = 1280 ms/image
Cine Magnetic Resonance Imaging

Conventional

- Prepare k-spaces ahead of time
- Instead of image by image, go line by line every repetition of the event
- Temporal resolution: limited by time required for one line (10 ms => 100 frames/s)
- Spatial resolution: limited by repetition
- Repetition triggered by some external signal or can be synchronized at post processing
- Challenges: human consistency, data synchronization

/i-a/
Cine MRI for Speech

Key Developments


– ECG (R peak) triggered acquisition: cue to subject as well as the MR system with 40 ms pause between each acquisition

– 200 repetitions acquired in 4 minutes of /pai/ and used for 3D reconstruction


– Noise bursts to trigger subject production and MRI acquisition


– Re-organization of k-space during post processing using speech and MRI noise recordings


– Acquire movies in each scan plane and do 3D reconstruction
Tagged Cine MRI for Speech

Key Developments

ECG triggered cine acquisition

Tags: sequence of RF and gradient pulses to saturate planes of magnetization (e.g., perpendicular to the imaging plane)

Different gridding schemes to capture internal and surface deformations

- Planar tagging
- Gridded tagging
- Checker-board tagging (MICSR)

Ref: http://www.speech.umaryland.edu (Vocal Tract Visualization Lab.)
Dynamic MRI

HOWEVER............
Toward real time acquisition for speech (circa 2004-)

Improving MRI temporal resolution
- A non 2D-FFT acquisition strategy (spiral k-space trajectory) on a GE Signa 1.5T CV/i scanner with a low-flip angle spiral gradient echo, 9-10 images/second
- Adapted pulse sequence originally developed for cardiac imaging.
- Effective reconstruction rates of 24-35 frames/second
  - sliding window reconstruction technique

First to use real-time MRI and synchronous noise-cancelled audio to understand vocal tract movements during natural speech production.

Data Collection Set up

MR technician

Two experimenters for:
  – subject consenting and instructing
  – interacting with subject (via intercom)

Audio acquisition person

Total run time:
  – 1 hour set up; 45 minute for subject prep and head coil and microphone placement
  – 1 hour run; 1 hour data backup

Processing time is significant
Real-time MRI sample clips

- Midsaggital (2D) view using 13-interleaf spiral pulse sequence
- 68-by-68 pixel (3mm-by-3mm) in-plane resolution
- 5mm slice thickness
- from: 11fps true rate/23fps sliding window reconstruction (2004) to up to 96 fps (2014)
- Synchronized audio recording
- Spatial resolution: 3mm/pixel
- Temporal resolution: 45ms/frame

Need to handle huge amounts of data!
Dynamic 3D visualization

HOW DO DIFFERENT TECHNIQUES COMPARE?

**EMA** (Wrench 2000)  
X-Ray Microbeam, XRMB  
(Westbury 1994)

**rt** Magnetic Resonance Imaging, MRI  
(Narayanan 2004)

Ultrasound  
(Stone 1980; Whalen 2005)

<table>
<thead>
<tr>
<th>Technique</th>
<th>Full mid-sagittal (or any section) view; 3D</th>
<th>Tongue (partial, surface view)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fleshpoints</strong></td>
<td>Full mid-sagittal (or any section) view; 3D</td>
<td>Minimally invasive</td>
</tr>
<tr>
<td><strong>Invasive</strong></td>
<td>Non-invasive</td>
<td>Portable, Easy</td>
</tr>
<tr>
<td><strong>Cumbersome</strong></td>
<td>Non-invasive</td>
<td></td>
</tr>
<tr>
<td>~100-500 Hz</td>
<td>~20-30 Hz</td>
<td>~50-300 Hz</td>
</tr>
</tbody>
</table>

SOME DATA SUITABLE FOR “TECHNOLOGY” STUDIES

XRMB (Univ. of Wisconsin) [1]

www.uni-jena.de/~x1siad/uwxrmbdb.html

- 32 F, 25 M; 118 different tasks incl. read sentences, paragraphs

MOCHA-TIMIT (Univ. of Edinburgh)[2]

http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

- One male and one female subject, each reading 460 TIMIT utterances
- Pre-processing of seven articulatory trajectories (500Hz)

EMA Database@MURI (Univ. of Southern California)[3]

http://sail.usc.edu/data.php

- One male American; Spontaneous conversations of 14 sessions (each ~5min)
- Pre-processing of six articulatory trajectories (200Hz)


The mngu0 database
http://www.mngu0.org

• EMA, MRI, Dental Casts Audio (from Edinburgh, LMU, Saarland)
  – EMA: Articulators: Upper and lower lips, jaw, and three tongue points; 1,300 utterances
  – MRI: 3D volume 13 vowels, 16 consonants & Midsagittal “dynamic” scans CVCs, (C=16,V=3)
**USC-TIMIT: A MULTIMODAL ARTICULATORY DATA CORPUS FOR SPEECH RESEARCH**

- 10 American English talkers (5M, 5F).
- Real time MRI (5 speakers also with EMA) and synchronized audio.
- 460 sentences each (>20 minutes)
- Freely available for speech research.


WEB-LINK (with download info):
http://sail.usc.edu/span/usc-timit/
SAIL homepage: http://sail.usc.edu
Some USC-TIMIT examples

Dynamic MRI

OK, now what...........?
Allows exploration of novel data-driven and hybrid knowledge-inspired approaches & modeling
Rest of the lecture

• Deriving articulatory representations
  - Direct methods
    Raw measures
    Derived task measures
  - Inverse methods

• Some case studies
  - Back to basics: learning from data
  - Vocal tract morphology
  - Articulatory setting
  - Relation between articulatory & acoustic representations
  - ASR and Speaker Verification
DIRECT MEASURES FROM DATA
**ARTICULATORY PHONOLOGY-BASED “GESTURES”**

**Gestural hypothesis:**
Act of speaking can be decomposed into atomic units of action, or gestures.

Gestures are dynamically controlled constriction actions of distinct vocal tract organs. (e.g., lips, tongue tip, tongue body, velum, glottis)

Gestural scores (Browman and Goldstein, 1992, 1995) represent latent activation intervals for dynamical systems controlling constrictions.

ARTICULATORY POSTURE & CONSTRICTION TASK VARIABLES

These feature sets are useful for modeling speech production dynamics


Image Analysis

Edge Detection
Canny (1991)
Bresch (2009)

Region-of-Interest
Lammert (2010)
Lammert (2013)

Grid-Based
Proctor (2010)
MR Image processing and analysis
A variety of needs and possibilities..

• Automatically track vocal tract outlines

• Automatic extraction of geometric features
  – Specific constrictions (e.g. velum, pharyngeal wall)
  – Specific contours (e.g. tongue)
  – Vocal tract aperture function

• **Derive other relevant data**
  – Aperture/Area function
  – Tract variable time series
  – Vocal tract resonance frequencies
  – Shapes

• **Novel direct image analysis and modeling**
“Model based” MR image analysis

• Commonly, define controlled articulators
  – larynx, epiglottis, tongue, lower lip
  – pharyngeal wall, glottis
  – velum, hard palate, upper lip

• Image processing tasks
  – find air-tissue boundaries along vocal tract
  – identify articulators
  – automated processing

• Technical Challenges of MR image analysis: noisy images
  – motion artifacts
  – off-resonance blur
  – spatially varying attenuation through non-uniform coil sensitivity pattern
  – subject dependent morphology of vocal tract
  – complicated deformation spaces, difficult to parameterize
  – few anatomical landmarks for registration
First define a contour model segmentation *manually*:
: each articulator in a different color

Now *hierarchically optimize* the model fit to the image in the Fourier domain using *gradient descent*

Algorithm Implementation

Operation in spatial frequency domain:
• native domain of MRI data
• difference image energy can be expressed in Fourier domain
• gradient of objective function can be expressed in closed form in Fourier domain
Algorithm Implementation

• Non-convexity
  – addressed through 4-level hierarchical optimization scheme
  – design of alternate gradient vector flow
  – points of anatomically connected section are moved in concord

Example: vowel [a] extracted from a read speech sequence
Sample Segmented Images

English

German
Problem: Cross-distance extraction from rtMRI data without appealing to contour tracking

Approach: Region-of-interest analysis, based on mean pixel intensity in pre-defined regions

Findings: High accuracy, low computational cost


(Semi-) automatic and robust parametrization of rtMRI data

Magnetic resonance image enhancement
Lips and larynx tracking
Segmentation and distance function computation

Jangwon Kim et al., “Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data”, in ISSP, 2014
What can we do with all these data?
A few example analysis studies

- Quasi-static Segmental: English fricative tongue shaping
- Dynamic Segmental: English nasals velum-oral coordination
- Dynamic Suprasegmental: Pauses
- Dynamic Paralinguistic: Affect
- Dynamic non-speech vocal production: singing

& some modeling and applications

- Vocal tract structure
- Articulatory setting
- Articulatory-acoustic relations
- Informing automatic speech recognition, speaker verification
Seek confirmatory/deeper/newer insights into well known questions in linguistics, speech science with traditional methods
Segmental speech characteristics

Tongue shaping of English sibilant fricatives /s/ and /sh/ in various vowel contexts

“Go pasop ok. Go pashop ok.”

There are more stimuli in corpus: “paseep”, “peesop”, “peeseep” etc.
Research studies – quasi-static

Tongue shaping of English sibilant fricatives /s/ and /sh/ in various vowel contexts

Some findings:

/s/ has a deeper tongue groove than /sh/.

Tongue surface for /sh/ is more parallel to the palate than for /s/.

Dynamic characteristics
coordination between adjacent segments and linguistic structure?

Velum-oral coordination of English nasals

Systematic timing differences between tongue
and velum constriction forming events?

Direct observation of tongue and velum tract
variables TTCD, VEL

Nasal position
Onset: /bow-know/, /toe-node/
Coda: /bone-oh/, /tone-ode/
Juncture geminate: /bone-know/, /tone-node/

❖ Data processing
  • Segment stimuli from carrier
  • Trace vocal tract
  • Measure VEL, TTCD constriction degree time series

❖ Define timing criteria
  • Time lag, e.g., w.r.t. 95% threshold
❖ Evaluate statistical significance of lag measures
Results

Velum-oral coordination of English nasals

• The velum opening lags behind tongue tip closure if the nasal is in onset position.

• Intergestural timing patterns sensitive to local stress context ==> Underlying timing specification that can yield flexibly

planning pauses in speech..?
Why are pauses informative?

- Aid or impair speech understanding
- Reflect the speech planning process

Hypothesis:

**Grammatical pauses are choreographed by a central cognitive planner**

- Planned pausing: slowing down of speech ‘clock’
- Unplanned pauses interfere with articulators reaching their targets

Mean gradient Energy 500 msec before & After Pause

- 7 subjects; 20-30s spontaneous speech
  - responses to questions like “tell me more about your favorite music”…
- Grammatical and ungrammatical pauses annotated and verified by a linguist
  - grammatical: Silent or filled pause that occurs at/between overt syntactic constituents
Are grammatical pauses are planned by a central ‘cognitive planner’?

2-way unbalanced parametric ANOVA
posthoc Tukey multi-comparison tests

Hypothesis is supported
• Interruption of speech flow
  • cognitively planned way
  • perturbation when this normal planning process fails
• ‘Recovering’ from unplanned pause that perturbs the linguistic structural integrity of the utterance

New technology to study articulation and acoustics jointly

Significant with p=0.01
Not significant p=0.01

going beyond linguistic details....

emotional speech production

Emotion effects at the Segment level: Vowel Triangle Example

First and second formant frequencies for the three vowels, /IY/, /AE/, /UW/ for various emotions.

Distinct constellation for different emotions: Emotions have different effects on different phonemes.

Notice that the low vowels /AE/ and /UW/ more affected by emotions than high vowel /IY/.

Example MRI Movies of Emotional Speech Production

neutral  angry  sad  happy

First two principal components of tongue shape contours

word “doctor” (subject AB)

Tongue tip positioning differences across emotions
Across emotions

• averaged articulatory trajectories are preserved, but
• movement ranges and velocities vary
• horizontal and vertical shifts of the tongue tip movements can be observed

Emotional Speech Production

Emotion, for a given linguistic context, modulates

• kinematics (i.e., **movement range and velocity**)
• constriction place (e.g., **tongue-tip positioning**).

• Speakers seem to **exploit possible articulatory freedom** for emotional contrast
  • within allowed variability in a given linguistic context

• Emotion expression does not affect the non-linear timing pattern variations much
  • localized variation in articulatory timing after duration normalization

Implications for information encoding model for speech communication


what are the soprano singers really doing..?

“The Soprano Challenge” : Sung vowels

- “Resonance tuning and vocal tract shaping in soprano singing”
- How do opera singers tune the vocal tract resonances to match the pitch while singing? (Remember they have to be loud!)
- Focus on vowel sounds: “LA”, “LE”, “LI”, ...
- Derive vocal tract resonance frequencies from images instead of audio, which would be difficult at high F0
Image analysis examples /a/

- spoken
- F0 = 233Hz
- Note 1
- F0 = 349Hz
- Note 5
- F0 = 622Hz
- Note 11
- F0 = 932Hz
- Note 15
Image analysis examples /i/

- spoken
- $F_0 = 233\text{Hz}$
- Note 1
- $F_0 = 349\text{Hz}$
- Note 5
- $F_0 = 622\text{Hz}$
- Note 11
- $F_0 = 932\text{Hz}$
- Note 15
Vocal tract shapes in comparison

- /a/
- /e/
- /i/
- /o/
- /u/

F0 = 233 Hz
F0 = 932 Hz
Sung vowel: Resonance tuning

Resonance tuning can be shown for vowels with low F1.
A few example analysis studies

- Quasi-static Segmental: English fricative tongue shaping
- Dynamic Segmental: English nasals velum-oral coordination
- Dynamic Suprasegmental: Pauses
- Dynamic Paralinguistic: Affect
- Dynamic non-speech vocal production: singing

& some modeling and applications

- Vocal tract structure
- Articulatory setting
- Articulatory-acoustic relations
- Informing automatic speech recognition, speaker verification
INTERSPEAKER VARIABILITY: PHYSICAL MORPHOLOGY

- Confined articulatory environment
- Variability across speakers
- Highly articulated, layered system
- Reflected in acoustical properties

Midsagittal view of vocal tract: 4 different subjects
Why is morphological structure relevant?

Palate Shape – Principal Components

46% of Variance
Component 1: concavity

30% of Variance
Component 2: anteriority

10% of Variance
Component 3: sharpness

Concavity: impact on F1 and F2; Anteriority: impact on F2 only; Sharpness: marginal

Experiments on estimating some of these shape details from acoustics...

Ming Li, Adam Lammert, Jangwon Kim, Prasanta Ghosh and Shrikanth Narayanan. Automatic Classification of Palatal and Pharyngeal Wall Shape Categories from Speech Acoustics and Inverted Articulatory Signals. SPASR, 2013
Can we estimate inherent morphological characteristics from acoustic data? (e.g., vt length, palate shape)

Challenge: ignore effects of controller
Inversion: Introduction

Motor Controller

Vocal Tract Structure

Speech Articulation

Speech Signal

Vocal Tract Shape

compensation

residual difference?

speaker-specific

invert?
Inversion: Palatal Concavity

Motor Controller

Binary Classification
concave or flat palate?

Speech Articulation

Vocal Tract Shape

Speaker ID Features
- MFCC
- Open-smile
- GMM UWPP
- Inv. artic. features

Inversion Accuracy: 63% - 71%

Inversion: Vocal Tract Length

Motor Controller → Speech Articulation

Vocal Tract Structure → Vocal Tract Shape → Speech Signal

compensation limited
... speaker-specific acoustics

invert vocal tract length?
Variation in Vocal Tract Length, Acoustics

Vorperian (2009)

Vocal Tract Length (cm) vs. Age in Years

Vocal Tract Length (cm) vs. Frequency (Hz)

F1, F2, F3
Vowel Variation & Vocal Tract Length

F1 (Hz) vs. F2 (Hz)

- /i/ (13.2 cm vt)
- /u/ (16.4 cm vt)
- /a/
VT Length Estimation: Model

\[ L = \frac{c}{4\Phi} \]

\[ \hat{\Phi} = \frac{F_n}{2n - 1}, \quad n = 1, 2, 3, \ldots \]

Model:

\[ \hat{\Phi} = \frac{\beta_1 F_1}{1} + \frac{\beta_2 F_2}{3} + \frac{\beta_3 F_3}{5} + \ldots + \frac{\beta_m F_m}{2m - 1} \]
VT Length Estimation: Design

\[ \hat{L} = \frac{c}{4\hat{\Phi}} \]

\[ \hat{\Phi} = \frac{\beta_1 F_1}{1} + \frac{\beta_2 F_2}{3} + \frac{\beta_3 F_3}{5} + \ldots + \frac{\beta_m F_m}{2m - 1} \]

Wakita (1977)
\[ \begin{align*}
\beta_1 &= 0 \\
\beta_2 &= 0 \\
\beta_3 &= 0.5 \\
\beta_4 &= 0.5
\end{align*} \]

Fitch (1997)
\[ \begin{align*}
\beta_1 &= -0.167 \\
\beta_2 &= 0 \\
\beta_3 &= 0 \\
\beta_4 &= 1.167
\end{align*} \]

Proposed
Determine \( \beta \) via linear regression
ESTIMATING VT LENGTH FROM ACOUSTICS

- 5 SPAN-TIMIT Subjects: real time MRI data
- Median estimated value (30 sec. read speech)

Could be used to account for speaker differences: new strategies for Vocal Tract Length Normalization
production information
for automatic speech recognition?

some results


TOWARD SPEAKER-INDEPENDENT ASR

**IDEA:** Estimate a mapping from the given speaker of interest to an “exemplar” speaker.

Ghosh and Narayanan (2011)

Role of speech production in speech recognition: the proposed approach

The speech production knowledge of an exemplary subject is used to estimate the articulatory features from talker’s speech

Augment acoustic features for speech recognition
Role of speech production in speech recognition: the proposed approach

Typical Articulatory Features – Raw data e.g., Electromagnetic Articulography (EMA) or Tract Variables (TV)

Typical Acoustic Features – MFCCs

Talker

Exemplary Subject (Exemplar)

Acoustic-to-articulatory mapping

Estimated Articulatory Features

Acoustic Articulatory Model

Recognized Phoneme Sequence

Speech Production Knowledge

Acoustic Features

Acoustic Feature Extraction

MFCC

EMA or TV
Role of speech production in speech recognition: the proposed approach

- Articulatory features need to be estimated for any arbitrary talker

Exemplar-specific
Talker-independent
Acoustic-to-articulatory inversion

Speech Production Knowledge

Recognized Phoneme Sequence

Acoustic Articulatory Model

Acoustic Feature Extraction

Estimated Articulatory Features

Acoustic Features

Acoustic-to-articulatory mapping

Exemplary Subject (Exemplar)

EMA or TV

MFCC

Talker

Exemplar-specific
Talker-independent
Acoustic-to-articulatory inversion
Exemplar-specific talker-independent acoustic-to-articulatory inversion

A Female Exemplar

Training

A Male Talker

Test

Listener-specific articulator-to-acoustic mapping

Articulatory feature vectors

Acoustic feature vectors

$\mathbf{x}_i \rightarrow \mathbf{z}_i \rightarrow \mathbf{x}^*_n \rightarrow \mathbf{u}_n$
Some notation

\( x_i \) \hspace{1cm} \text{Train Articulatory Feature vector, } 1 \leq i \leq T

\( Z_i \) \hspace{1cm} \text{Train Acoustic Feature vector, } 1 \leq i \leq T

\( u_n \) \hspace{1cm} \text{Test Acoustic Feature vector, } 1 \leq n \leq N

\( x^*_n \) \hspace{1cm} \text{Target Articulatory Feature vector, } 1 \leq n \leq N

\( \eta^{l,j}_n \) \hspace{1cm} L \text{ possible values, } 1 \leq n \leq L, \text{ of } j^{th} \text{ Articulatory Feature vector trajectory at frame } n

\( p^l_n \) \hspace{1cm} \text{Probability that } \eta^{l}_n \text{ is the value of articulatory position given } u_n \text{ is the test acoustic feature}

Both determined from \((z_i, x_i), 1 \leq i \leq T, \text{Training data}\)

Best articulatory sequence \( x^*_n \) obtained through optimizing a criterion:
1. Smoothness term
2. Proximity term: such that test and train acoustic features are close (e.g., Euclidean sense)
Generalized Smoothness Criterion (GSC) for inversion

Consider only the $j$-th articulatory trajectory $x^j_n$ and find its possible values $\gamma_{n,i,j}$ at each frame with their probability $p^i_n$.

And then fit the best trajectory which is smooth to the required degree.

And repeat these steps for all articulatory feature trajectories.
Generalized Smoothness Criterion (GSC) for inversion

\[ \{ \chi_n^{j*} ; 1 \leq n \leq N \} = \arg \min_{\{ \chi_n \}} \left\{ \sum_n \left( \chi_n^j \cdot h^j [n] \right) + C_j \sum_n \sum_l \left( \chi_n^j - \eta_{n,l}^{l,j} \right) p_n^l \right\} \]

Articulator specific smoothness

Data proximity term

Trade-off parameter

\[ x^{j*} = \begin{bmatrix} x_1^{j*} & \cdots & x_N^{j*} \end{bmatrix}^T = \left( R^j + C_j I \right)^{-1} d^j \]

\[ R^j = \{ R_{pq}^j \} = \{ R^j (p - q) \} = \sum_n h^j [n - p] h^j [n - q] \]

\[ d^j = \begin{bmatrix} C_j \sum_l \eta_{1,l}^{l,j} p_1^l & \cdots & C_j \sum_l \eta_{N,l}^{l,j} p_N^l \end{bmatrix}^T \]

- The solution can be derived recursively (over \( n \)) without any performance loss; hence, in practice no matrix inversion is required

Exemplars considered in preliminary work

Male (MOCHA-TIMIT\(^1\))
Exemplar 1

Female (MOCHA-TIMIT\(^1\))
Exemplar 2

Male (MURI\(^2\))
Exemplar 3


Exemplar-specific talker-independent acoustic-to-articulatory inversion using GSC

We observe a performance drop in accuracy for estimated articulatory features, but with the advantage of not requiring production info from that speaker.
Listener-specific approach – comparison to other production-oriented approaches to ASR

Feature Detection: Integration with HMM
[Kirchhoff '98 '99 '02]

Articulatory configuration as HMM states
[Richardson '98], [Deng '02]

Dynamic Bayesian Network
[Blimes '99], [Zweig '02], [Ostendorf '00]

Dynamical Systems
[Bakis '91], [Richards '99]

MISCELLANEOUS
[MO_MALCOM 'Hodgen '00]

Articulatory Data-driven Approach
[Frankel '01]

Speech producing HMM for recognition [Tokuda '03]

• Real articulator data driven
• Can be extended for any corpus and any listener

Real articulator data driven
Can be extended for any corpus and any listener

Position
Transition
Phone
Articulators: Tongue, Lips
Acoustic Observation

101
Using MFCC + estimated articulatory features (speaker-independent inversion) improves speaker recognition performance!

Ming Li, Jangwon Kim, Adam Lammert, Prasanta Ghosh, Vikram Ramanarayanan and Shrikanth Narayanan. Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Computer, Speech, and Language. 2015
Can new data and new computational techniques lend new scientific insights?
Exploring extraction of primitives through cNMF and validation through phone discrimination


ARTICULATORY PRIMITIVES

A set of **time-varying functional units** ("synergies") or basis functions, **weighted combinations** of which can be used to represent any movement of articulators in the vocal tract.
\[ \mathbf{V} = \begin{bmatrix} I_1 & I_2 & \ldots & I_N \end{bmatrix} \in \mathbb{R}^{M \times N} \]

\[ \mathbf{W}(t) \in \mathbb{R}_{\geq 0}^{M \times K}, \quad K \leq M \]

\[ \mathbf{H} \in \mathbb{R}_{\geq 0}^{K \times N} \]

\[ \min_{\mathbf{W}, \mathbf{H}} \| \mathbf{V} - \sum_{t=0}^{T-1} \mathbf{W}(t) \cdot \mathbf{H}^t \| \]

where \( h_i \) is the \( i \)-th row of \( \mathbf{H} \) and \( 0 \leq S_h \leq 1 \) is user-defined.


These features perform competently when compared to MFCC/EMA features with minimal loss in information content!

Spatio-temporal movement primitives can be used as features in phone classification experiments.

<table>
<thead>
<tr>
<th>Feature set $\mathcal{X}$</th>
<th>Class. Acc. (%)</th>
<th>$H(\mathcal{X})$</th>
<th>$I(\mathcal{X}; \mathcal{L})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>71%</td>
<td>6.9</td>
<td>1.68</td>
</tr>
<tr>
<td>Raw EMA pellets</td>
<td>61.78%</td>
<td>6.9</td>
<td>1.59</td>
</tr>
<tr>
<td>Primitive activations</td>
<td>80.59%</td>
<td>6.5</td>
<td>1.63</td>
</tr>
<tr>
<td>Phone labels $\mathcal{L}$</td>
<td>100%</td>
<td>4.9</td>
<td>4.9</td>
</tr>
</tbody>
</table>

The measured EMA trajectories are used as input to the classification module.
Challenges & Opportunities

- **Data acquisition**
  - Improving quality of any single modality especially imaging ones
  - Putting different pieces of the data puzzle together ("co-registration")
  - The scale of data operation still very small

- **Seeking meaningful representations**
  - Interpretability
  - Modeling specific linguistic or indexical construct

- **Modeling**
  - Reconciling, and revising, theoretical accounts from planning and control to the execution
  - Informing technology applications: “Meeting the bar” set for technological translation by demonstrating the advantage
Accelerated acquisition enables imaging fast articulatory movements.
Custom upper-airway coil

- We use custom coils over commercial coils (developed for other body parts)
  - Superior SNR in upper-airway regions
  - Utilized (later) towards accelerated imaging

<table>
<thead>
<tr>
<th>ROI</th>
<th>Upper lip</th>
<th>Lower lip</th>
<th>Front tongue</th>
<th>Middle tongue</th>
<th>Back tongue</th>
<th>Velum</th>
<th>Pharyngeal wall</th>
<th>Epiglottis</th>
<th>Glottis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative SNR gain (SNR_{UA}/SNR_{Brain})</td>
<td>3</td>
<td>4.3</td>
<td>5.4</td>
<td>4.5</td>
<td>4.6</td>
<td>2.2</td>
<td>2.6</td>
<td>5.9</td>
<td>5.2</td>
</tr>
</tbody>
</table>
Accelerated RT MRI of speech

- Regularized SENSE parallel imaging with temporal finite difference sparsity

\[ \min_f \left( \| A(f) - b \|^2_2 + \lambda \| \nabla_t(f) \|_1 \right) \]

- Data consistency
- Temporal reg.

\[ A \] - coil sensitivity encoding + NUFFT along GA spiral

\[ \nabla_t \] - temporal finite difference

---


Seeing fast speech @ 12 ms/frame

Repetitions of "one-two-three-four-five" at normal pace followed by rapid pace

2.4 mm²

zoomed in time profiles

superior temporal fidelity
Clinical Applications


Concluding Remarks

• **Data is integral to advancing speech communication research**
  • Vocal tract information provides a crucial piece of the puzzle
  • Need to gather and integrate multiple, disparate sources of information toward getting a more complete picture of speech production

• **The problem is highly challenging**
  • Technological, computational as well as conceptual/theoretical challenges
  • Potential for applications including machine speech recognition, speaker modeling and synthesis

• **Acquiring, interpreting and utilizing speech production information is an ongoing interdisciplinary scientific endeavor**
Thanks to USC SPAN Team/Alums
sail.usc.edu/span

Dani Byrd  Krishna Nayak  Sungbok Lee  Louis Goldstein  Asterios Toutios

Adam Lammert  Vikram Ramanarayanan  Jangwon Kim  Yinghua Zhu  Christina Hagedorn

Nassos Katsamanis  Mike Proctor  Yoon Kim  Erik Bresch  Prasanta Ghosh
ACKNOWLEDGEMENTS

- NIH Grants DC007124 & DC03172
- NSF, ONR, DoJ
- USC Imaging Sciences Center
- LAC-USC Hospital
- USC Center for High Performance Computing

Papers, Videos, Teaching resources

http://sail.usc.edu/span


• Ming Li, Jangwon Kim, Adam Lammert, Prasanta Ghosh, Vikram Ramanarayanan and Shrikanth Narayanan. Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Computer, Speech, and Language. 2015 (Also, Interspeech, 2013).

DATABASES/WEBSITES with MULTIMEDIA RESOURCES

• USC TIMT CORPUS


http://sail.usc.edu/span/usctimit/

• USC EMO MRI CORPUS


• USC EMA CORPUS

Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh and Shrikanth S. Narayanan, An articulatory study of emotional speech production, in Proceedings of InterSpeech, pages 497-500, 2005

http://sail.usc.edu/ema_web/index.html

MUSIC


http://sail.usc.edu/span/beatboxing/


http://sail.usc.edu/span/videos/USC-Soprano-AveMaria.mov