Articulatory Phonology:
Looking backward and forward
Looking back...

- What were the issues that gave rise to this approach?
- What were the key ideas that Articulatory Phonology developed to address these issues?
- In the years since, what have we learned about phonology and speech using these key ideas?
- Looking forward, what might we still hope to learn using these ideas.

Cathe Browman ca, 1985
The Problem:
Apparently incompatible descriptions of speech

“Jane may earn more money by working hard”

- Phonological
  - sequence of discrete symbols from a small inventory that recombine to form different words
- Physical
  - continuous, context-dependent variation in many articulatory, aerodynamic, acoustic parameters
The problem... Fowler, 1976; 1980

<table>
<thead>
<tr>
<th>Phonological Units</th>
<th>Physical Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete</td>
<td>Continuous</td>
</tr>
<tr>
<td>Time-invariant</td>
<td>Time-varying</td>
</tr>
<tr>
<td>Context-independent</td>
<td>Context-dependent</td>
</tr>
<tr>
<td>Low Dimensionality</td>
<td>High Dimensionality</td>
</tr>
</tbody>
</table>

- Historically (e.g. Hockett, 1960) the incompatibility is “resolved” by separating cognitive and physical descriptions.
- But this just pushes back the problem.
Articulatory Phonology: Key ideas of resolution

- Dynamical systems representation
- Articulatory Synergy representation
- Temporal overlap of subsegmental units

Gestures

<table>
<thead>
<tr>
<th>Phonological Units</th>
<th>Physical Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete</td>
<td>Continuous</td>
</tr>
<tr>
<td>Time-invariant</td>
<td>Time-varying</td>
</tr>
<tr>
<td>Low Dimensionality</td>
<td>High Dimensionality</td>
</tr>
<tr>
<td>Context-independent</td>
<td>Context-dependent</td>
</tr>
</tbody>
</table>
Invariant subsegmental articulatory units (eventually gestures) with temporal extent may exhibit various temporal alignments to each other and to the units of neighboring segments, including partial and complete overlap.

“mis”-alignment of units can have phonological consequences.

- LINGUA
  - Demi-syllable text-to-speech system
  - but each feature had its own temporal extent and rules could specify their alignment

Autosegmental Phonology
Goldsmith, Leben, Anderson

Fujimura

Subsegmental Units: What has been learned?

- Subsegmental units (gestures) do not float freely but are **coordinated** to one another in phonologically-specific ways.

- Coordination may involve synchrony (at onsets) or sequentiality, and more...

- Organization of gestures into segments is **not** a good predictor of coordination type (but syllable structure is):
  - Gestures belonging to a sequence of segments may be coordinated synchronously (e.g. onset consonant and vowel).
  - Gestures belonging to a single segment may be coordinated sequentially (e.g., velic and oral gestures of a nasal in coda).
Variation in temporal overlap

- Gestures at the boundaries of words (or syllables) can overlap at relatively fast speech rates,

- Overlap can produce perceived assimilations (Byrd, 1992) or deletions, even when produced gestures are unaltered.

- Important role of prosody in “segmental” alternations. from Tiede et al. 2004.
Variation in temporal overlap

- Gestures at the boundaries of words (or syllables) can overlap at relatively fast speech rates,

- Overlap can produce perceived assimilations (Byrd, 1992) or deletions, even when produced gestures are unaltered.

- Important role of prosody in “segmental” alternations. from Tiede et al. 2004.
Modeling variation due to prosody: “told before”

- Phrase was generated automatically by computational model (TaDA), then boundary gesture slowing effects were added, using π-gesture (Byrd & Saltzman, 2003)
- [d] release emerged automatically at boundary, due to decrease in overlap of Tongue Tip and Lip Gestures.

Only 2/38 Transcribers hear boundary

11/38 Transcribers hear boundary

TaDA:
Subsegmental Units: What has been learned?

- Early hypothesis (Browman & Goldstein, 1990)
  - All post-lexical assimilatory alternations arise from systematic changes in overlap or reduction in gesture magnitude.
  - This hypothesis was not supported.
  - E.g., place assimilation in Spanish (Honorof, 1998)
    
    Digan “paja” alto.
    Digan “caja” alto.

Final /n/ assimilates to following labial or dorsal.
Coronal gesture is fully deleted and resulting labial or dorsal is longer than it is in control utterances with no final /n/. (e.g., Diga “paja” alto).
Coordination relations in grammar (Gafos, 2002; 2019)

- In Moroccan Arabic, coordination relation between consonant gestures plays a critical role in morphophonology.
- In stem-affix junctures, sequences of homorganic consonants are coordinated with partial overlap such that there is no release.
- In morphological templates, homorganic consonants are coordinated with no overlap such that there is a release between them (C↔C).
- That constraint in templates conflicts with the preferred coordination relation in templates (CoC), as exhibited in heterorganic sequences.
- Conflict plays out in a variety of ways in template satisfaction.
Subsegmental timing and overlap: Going forward

• How can distinct, abstract coordination patterns be formally characterized, within a dynamical model?

• Contrasting coordination patterns in Moroccan Arabic

• Cross-language typology in degree of overlap of C sequences across syllables (Zsiga, 2000 Kochetov et al., 2007; Yanagawa, 2006)
  • Russian, Georgian little or no overlap \( C \leftrightarrow C \)
  • German, English partial overlap \( C \cap C \)
  • Cantonese near complete overlap

• Within-word vs. across-word overlap. prosodic boundaries?

• Across languages \( C_1 C_2 \) sequences in which \( C_1 \) is anterior to \( C_2 \) are produced with more overlap than when \( C_2 \) anterior to \( C_1 \) (Chitoran et al., 2002)

• Why? And does the answer provide inside into the more fundamental spatiotemporal mechanisms of sequencing?

• Segment-internal gestural coordination & phonotactics e.g., English rimes with liquids (Walker & Proctor, sub)

• Subsegmental timing and Q-theory (Inkelas & Shih, 2016)
Articulatory Phonology: Key ideas of resolution

- Dynamical systems representation
- Articulatory Synergy representation
- Temporal overlap of subsegmental units

Gestures

<table>
<thead>
<tr>
<th>Phonological Units</th>
<th>Physical Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete</td>
<td>Continuous</td>
</tr>
<tr>
<td>Time-invariant</td>
<td>Time-varying</td>
</tr>
<tr>
<td>Low Dimensionality</td>
<td>High Dimensionality</td>
</tr>
<tr>
<td>Context-independent</td>
<td>Context-dependent</td>
</tr>
</tbody>
</table>
Dynamical Systems

- Two parts of system:
  - **State**: quantity whose change over time (t) or space (x) or both is being described
  - **Law**: local relation between neighboring states

Law holds everywhere in space or time.
Law is fixed while state is changing.

If phonological units are dynamical laws, time-varying and time-invariant are lawfully related.

---

Coarticulation and theories of extrinsic timing control
*Journal of Phonetics*, Vol. 8 pp. 113-133

Implications for speech production of a general theory of action

Kelso, J. A. S., Tuller, B., & Harris, K. S. (1983)
A 'dynamic pattern' perspective on the control and coordination of movement
In P. MacNeilage (Ed.), *The production of speech*, New York: Springer-Verlag. 7,
Context-independence and discreteness

• **Context-dependence:**
  Same context-independent Law will result in different context-dependent, continuous time functions as a function of initial conditions.

• **Discrete Attractors:**

• **All** Initial conditions result in same value after a long enough time.

• The system is seeking a goal: stable equilibrium or an attractor.

• Time function is continuous, but attractor is discrete.

• Same Law always has the same attractor, though it may fail to be reached.

\[ \Delta x = -0.5x \]
Dynamical system as model of phonological units

• Fowler (1976, 1980) was the first to see that defining phonological units as dynamical system could solve the problem of incompatible descriptions of phonology and phonetics.

• Phonological units are laws with fixed parameter values that are context-independent.

• The phonetic state \((x)\) corresponding to a phonological unit evolves over time in a continuous, context-dependent way.

• Solution is non-dualistic. Phonology and phonetics require no ‘interface’

• Phonological constants and time-varying phonetic quantities are in same equation.

\[
\Delta x = -kx + C
\]

- **Law**
  - \(C\) / \(k\)
- **Goal**
  - Stiffness: Rate of goal attainment

\(k\)
Dynamics & phonology: what have we learned

- **Dynamics of gestures and their real-time production:**
  - Adequate reproduction of intelligible speech that mimics signature properties of speech is possible modeling gestures as critically damped linear second order dynamics (task dynamics).
  - Some allophony can be modeled through undershoot: gestural activation < time to reach target (given k) spirantization in Spanish (Parrell, 2011; 2018)
  - Non-linear spring (‘soft’) may model kinematics better than linear system (Sorensen & Gafos, 2016)

- **Dynamics of gestural planning:** Nonlinear dynamical systems with multiple attractors
  - Coupled planning oscillators represented in coupling graph can account for robust observations of inter-gestural timing as a function of syllable structure and also variability in relative timing (Nam et al., 2009)
    - Onsets gestures and nucleus are coupled in-phase.
    - Nucleus and coda gestures are coupled anti-phase.
    - Multiple onset gestures are competitively coupled anti-phase, giving rise to C-Center effect in complex onsets, that can be used as a diagnostic of a complex onset.
  - Competitive queuing model of gestural sequencing (through selection) in combination with a phasing model for selected units can account for a variety of aspects of phonological structure, e.g., moraicity (Tilsen, 2016)
  - Phase transitions between attractors can emerge as control parameters vary
    - gestural intrusion speech errors (Goldstein et al., 2007)
    - discrete alternations, pre-aspiration—> post-aspiration in WAS (Parrell, 2012)
    - Re-syllabification (deJong, 2001)
Dynamics & phonology: what have we learned

- *Grammar dynamics*: shifts between attractors as a function of biasing parameters account for both discrete and continuous aspects of phonological alternations.
  - Harmony (Gafos & Benus, 2006)
  - Weight-to-stress (Goldsmith)
Phonological categories (and sequences) in a dynamical grammar

• Phonological units can be represented as dynamical systems in the state space of goal (task) variables.

• These systems have attractor states, determined by the $x_0$ of the dynamical system.

• But there is quantitative variability in the value of $x_0$ associated with a given phonological unit (e.g., gesture), some of which is not accounted for by gestural overlap.

• stochastic
• prosody and rate context
• other context?

• Also qualitative (category-changing) variability —alternations
Phonological categories (and sequences) in a dynamical grammar

• In a traditional grammar, the categories are purely symbolic, and not intrinsically connected to phonetic substance, so quantitative variability in $x_0$ does not undermine their belonging to the same category/unit.

• But in a dynamical grammar, they are intrinsically related, so what identifies the distinct $x_0$ as belonging to the same unit?
Dynamical Constraints (Gafos & Benus, 2006)

- Dynamical systems that model the preferences of the grammar to select certain phonetic states (e.g. $x_0$)

- A simple first-order system exhibits an attractor along some continuous state variable $x$, that represents a task variable, e.g. Lip Aperture.

- All trajectories converge at $LA = -2$.

- IDEA: The grammar produces only one stable value for this unit, no matter what kind of “input” the grammar begins with.

- ALTERNATE IDEA: Planning process derives only one goal for the unit, regardless of how planning is initiated.

\[
\frac{dx}{dt} = -k(x - x_0)
\]

\[
x_0 = -2
\]

\[
k = .95
\]
Potential Functions

• Attractors of a dynamical system of this type can be visualized by plotting its potential function \( V(x) \).

\[
V(x) = \frac{k(x - x_0)^2}{2}
\]

• The potential is obtained by integrating the differential equation.

\[
\frac{dx}{dt} = -k(x - x_0)
\]

• So therefore, the change of \( x \) in time is equal to the negative slope of the potential function.

\[
\frac{dx}{dt} = -\frac{d(V(x))}{dx}
\]

• Imagine dropping a ball on the potential function landscape.

- Where the slope is non-zero, there will be change in state towards the attractor.
- At the attractor, the slope is zero and there will be no change.

\( x_0 = -2 \quad k = .95 \)

\[ x = \text{Lip Aperture} \]
Stochastic Variation

• Noise term can be added to differential equation, which results in a distribution of final values of $x$ when differential equation is simulated multiple times.

• The result is a distribution of values that can be compared against observed distributions.

$$\frac{dx}{dy} = -k(x - x_0) + \zeta$$

$x = \text{Lip Aperture}$
Contextual Variation: Spanish stops
(Parrell, 2011)

- Spanish voiced stops vary between stops and approximants.

  - variable, but roughly: stops phrase-initially approximants medially

- Parrell found that the CD (LA) of /b/ is dependent on the gesture duration

  - Longer duration at prosodic boundary gives rise to full closure (narrower constriction degree)

  - Closure is undershot when gestural duration is shorter.

- However, there are ranges of duration that show overlap of /p/ and /b/, but their LA do not overlap: /b/ < /p/

  - x0 of /b/ and /p/ must differ slightly
Modelling /b/-/p/ in TaDA

- x0 differs for /b/ and /p/.
- So lip closure gesture (phonological unit) has a context-dependent target value depending on the associated laryngeal gesture:
  - larynx lowering for /b/
  - glottal abduction for /p/
- How do we model the context dependence dynamically?
Contextual Variation in Dynamical Law

- Adding an additional constant parameter $C$ to law can shift the value that will show minimum potential (attractor).

\[
\frac{dx}{dy} = -k(x - x_0) + C + \zeta
\]

- Effect is to “tilt” the potential function

- Single law for lip closure gesture with contextually varying value of $C$

\[
x_0 = -2 \\
C = 0
\]

\[
x_0 = -2 \\
C = 4.5
\]

\[
x - x_0 = 1.5
\]
Dynamics: Going Forward

• Interaction of planning and production models
• How does feedback of various kinds modulate ongoing planning?
• Planning models are purely temporal, not spatiotemporal
• Spatial properties of gestures modulate timing: Onset-nucleus gestures are not synchronous when the nucleus is a syllabic consonant. (Pouplier & Benus, 2011)
• Role of sonority in syllable structure
  • Dynamical syllabification model of Goldsmith (Goldsmith & Larson, 1990)
  • How can this be integrated with model of timing?
Articulatory Phonology: Key ideas of resolution

- Dynamical systems representation
- Articulatory Synergy representation
- Temporal overlap of subsegmental units

Gestures

<table>
<thead>
<tr>
<th>Phonological Units</th>
<th>Physical Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete</td>
<td>Continuous</td>
</tr>
<tr>
<td>Time-invariant</td>
<td>Time-varying</td>
</tr>
<tr>
<td>Low Dimensionality</td>
<td>High Dimensionality</td>
</tr>
<tr>
<td>Context-independent</td>
<td>Context-dependent</td>
</tr>
</tbody>
</table>
Articulatory Synergy Representation: Task Dynamics

• Performance of any skilled motor task requires cooperation of several independently moveable body parts, or articulators.
  • e.g., reaching for an object on a table

• There is large (possibly infinite) set of articulator postures that will achieve the task. This is sometimes called redundancy.

• When we learn to perform a task, we learn a pattern of dependency among the articulators specific to the task. This is called a synergy or a coordinative structure or more recently, a control law.

• The synergy allows the task to be performed in different ways
  • in different environmental contexts.
  • by different actors

Skilled Actions: A Task-Dynamic Approach
Psychological Review, Vol. 94 (1) pp. 84-106

A Dynamical Approach to Gestural Patterning in Speech Production
Task space for gestures

• Task dynamics and phonology:
  • Gestures are motor tasks performed by a set of articulators to achieve a goal.
  • Define dynamics at the level of the task: Dynamical system produces time-varying changes in the task variable that is then distributed across a set of articulators in a context-dependent way.

• In what space are the tasks of gestures specified?
  • Auditory properties
  • Acoustic properties
  • Aerodynamic properties
  • Vocal tract constrictions

Different gestures may have different task spaces
Hypothesis: Tasks are constrictions of 5 effector systems

<table>
<thead>
<tr>
<th>Task</th>
<th>Articulators</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP</td>
<td>upper &amp; lower lips, jaw</td>
</tr>
<tr>
<td>LA</td>
<td>upper &amp; lower lips, jaw</td>
</tr>
<tr>
<td>TTCL</td>
<td>tongue tip, tongue body, jaw</td>
</tr>
<tr>
<td>TTCD</td>
<td>tongue tip, tongue body, jaw</td>
</tr>
<tr>
<td>TBCL</td>
<td>tongue body, jaw</td>
</tr>
<tr>
<td>TBCD</td>
<td>tongue body, jaw</td>
</tr>
<tr>
<td>VEL</td>
<td>velum</td>
</tr>
<tr>
<td>GLO</td>
<td>glottis</td>
</tr>
</tbody>
</table>

ASY: Articulatory Synthesizer

Rubin, Baer & Mermelstein (1981)
Motivation for constriction tasks

(1) Major determinant of acoustic properties (formants)

(2) Utility:
Task dynamic model needs to know the forward map: q is a vector of articulators; x is a vector of task variables:

\[ x = k(q) \]

- And also the Jacobean:

\[
J(q) = \begin{pmatrix}
\frac{\partial x_1}{\partial q_1} & \cdots & \frac{\partial x_1}{\partial q_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial x_m}{\partial q_1} & \cdots & \frac{\partial x_m}{\partial q_n}
\end{pmatrix}
\]

- Given the simplified articulatory model (ASY), these can be calculated analytically from the geometry.

(3) Locality
Computational Model

• Because of lawful relation between phonological analysis as a structure of gestures and the resulting vocal tract movements and sounds, it was possible to test phonological hypotheses using quantitative data.

• But only with an explicit model…. GEST …TaDA
  • assimilations, deletions
  • temporal vs. spatial control (targetlessness)
  • syllable structure
Gestural score

• Input to task dynamics: activation interval of each gesture (Time during which its dynamical system governs the vocal tract).

• Labels in boxes represent the task goal dynamical parameter

“pan”
Representing Contrast in Gestural scores

- Presence or absence of a gesture employing a given constriction device (intrinsically discrete)
- Contrastive parameter values for a given task (for example a coronal stop differs from a coronal fricative in the Task Variable of Tongue Tip Constriction Degree)
- Temporal Organization
Locality of constriction tasks

• Constrictions produced by effectors are local within the vocal tract.

• Because the actions of constricting organs are partially independent of one another, multiple constrictions can be co-produced with minimal interaction, resulting in parallel transmission of information units (Liberman et al., 1967).

• In contrast, the acoustic properties of the vocal tract are global, they are dependent on all the constrictions that are active.
What have we learned about constrictions as tasks?

- Evidence supporting constrictions
  1. Constrictions (Location and Degree) are specified by formant frequencies and amplitudes (Iskarous, 2010)
  2. Constriction information contributes to (automatic) speech recognition (Mitra et al., 2011)
  3. Sites in the motor cortex during speaking code coordinated articulatory motion forming labial, coronal & dorsal constrictions (Chartier et al, 2019).

- Evidence supporting other task variables
  1. Role of auditory information in speech production (Houde & Jordan, 1998)
  2. Evidence for lowered F3 as goal of /r/ production (Nieto et al, 2005)
  3. Patterns of neural activation in the motor cortex during listening code auditory patterns, not constrictions. (Cheung et al., 2016).
1. Constrictions are specified by formant frequencies and amplitudes (Iskarous, 2010)

- If constriction gestures are phonological units, they should be engaged during speech perception and should therefore be recoverable from acoustics.

- It known that the relationship of fully specified vocal tract area functions to formant frequencies is many-to-one (Atal et al., 1976).

- However:
  - Constriction information is more constrained than arbitrary area functions
  - Taking formant bandwidths (or amplitudes) into account further constrains the solution.
Fourier decomposition of area functions  
(Iskarous, 2010, after Schroeder, Rice, Öhman)

- Area function of vocal tract can be decomposed into anti-symmetric and symmetric fourier components.
- The anti-symmetric components correspond to constrictions within tube that is closed on one end and open at the other.
- Symmetric components are relevant when lips are narrowed.

Anti-Symmetric

Symmetric

Fig. 1. Antisymmetric and symmetric basis functions of the discrete Fourier transform.
Recovery of constrictions

- Anti-symmetric components of area function are directly related to formant frequencies:

- Symmetric components can be estimated from formant amplitudes.

- Recovery of TBCD and TBCL from formant frequencies and amplitudes of data from XRMB:

![Graphs showing TBCD and TBCL estimates](image)

---

**Table 2**

<table>
<thead>
<tr>
<th></th>
<th>/i/</th>
<th>/I/</th>
<th>/e/</th>
<th>/æ/</th>
</tr>
</thead>
<tbody>
<tr>
<td>XRMB Measured</td>
<td>1.38</td>
<td>1.16</td>
<td>.92</td>
<td>.86</td>
</tr>
<tr>
<td>Formant-only</td>
<td>1.44</td>
<td>1.38</td>
<td>1.28</td>
<td>1.16</td>
</tr>
</tbody>
</table>

Table 3 presents the difference in means in estimating CDs for the front series.

**Table 3**

<table>
<thead>
<tr>
<th></th>
<th>/i/</th>
<th>/I/</th>
<th>/æ/</th>
</tr>
</thead>
<tbody>
<tr>
<td>XRMB Measured</td>
<td>1.28</td>
<td>1.16</td>
<td>.90</td>
</tr>
<tr>
<td>Formant-only</td>
<td>1.34</td>
<td>1.28</td>
<td>.95</td>
</tr>
</tbody>
</table>

**Table 4**

<table>
<thead>
<tr>
<th></th>
<th>/u/</th>
<th>/o/</th>
<th>/æ/</th>
</tr>
</thead>
<tbody>
<tr>
<td>XRMB Measured</td>
<td>2.08</td>
<td>1.75</td>
<td>.90</td>
</tr>
<tr>
<td>Formant-only</td>
<td>2.09</td>
<td>1.75</td>
<td>.90</td>
</tr>
</tbody>
</table>

**Notes**

- The lack of rounding for the vowel /æ/ is as unrounded as the front vowels, but they differ significantly from the acoustic estimate used here.
- The XRMB data (XRMB) presents 40 out of 72 possible pairs of measurements, with a significant difference marked by a black box around each pair.
- Anti-symmetric components, however, are pharyngeals. However, the CL vowel of /æ/ is estimated to be as unrounded as the front vowels, but they differ significantly from the acoustic estimate used here.
- The recovery of TBCD and TBCL from formant frequencies and amplitudes of data from XRMB is shown in the graphs.
2. Constriction information contributes to automatic speech recognition

- If gestures are phonological units that represent contrast, and they are defined using constriction tasks, recovering constriction time functions (Task Variables, or TVs) and Gestures (k, C) should aid ASR.

- Estimating TVs from acoustics should contribute to increased success in recognition in noise (robustness).

**Architecture of gesture-based ASR system**

Constriction TV estimation using ANN

• Model estimating TVs from MFCCs was developed for a synthetic data base of number strings using TaDA, since the TVs that actually generate the acoustics were known exactly.

• Single ANN trained for all TVs at once (addresses non-linearity of mapping) 3 hidden layers (150-100-150 nodes)

• Acoustic features: **MFCC**
  13 Mel-Frequency Cepstral Coefficients
  standard robust representation of vocal tract acoustics

• A **contextual** window was used (170msec)

• Estimated TVs were noisy → processed by Kalman smoothing

![Diagram of Artificial Neural Network](image)
Estimation of TVs from real data

- ANN models trained from on TaDA were applied to digit strings in the Aurora-2 database with car and subway noise.

![Graphs showing comparison between TaDA-generated and Aurora-2 data](image-url)
trajectories (Art-14) when used in addition to MFCCs, where the backend uses 3-mixture components per state. Fig. 12 clearly shows the superiority of TVs over Art-14 for improving the noise-robustness of a word-recognizer. Although Art-14 is found to improve the noise robustness over the MFCC baseline, it fails to perform as good as TVs.

C.3 Speech enhancement

This section examines how speech enhancement will interact with the use of TV estimates and MFCCs. We used the preprocessor based MPO-APP speech-enhancement architecture described in [64] to enhance the noisy speech signal from Aurora-2. Four different combinations of MFCC and TV estimates were obtained depending upon whether or not their input speech was enhanced. Fig. 13 presents the average word recognition accuracies obtained from these four different feature sets. Similar to the results MPO-APP [3] speech enhancement architecture was motivated by perceptual experiments.

The MFCC_MPO-APP and the TV_MPO-APP are the MFCCs and TVs that were obtained after performing MPO-APP enhancement of the speech signal.
Contribution of Constriction TVs

- **Substantial** improvement in word recognition in noise by adding TVs. Why??

- Acoustic signal is rich in information, much of it is not relevant to gestures or to word identification.

- When signal is clear, the richer acoustic signal (MFCC) performs much better than TVs alone.

- But when signal is noisy, the reduced information needs to project into subspace that is optimal for contrast. TVs alone actually perform a bit better than the MFCCs.

- Human listeners also use constriction information in making perceptual judgments under noisy conditions (Ausilio et al. 2009)

---

**Figure 11.** Average word recognition accuracy (average $d$ across all the noise types) for the baseline and TVs with different $\Delta$s.

**Figure 12.** Comparison of recognition accuracy of MFCC+TVs from the different word models to the baseline accuracy using MFCC only. Adding TVs to MFCCs resulted in significant improvement in the word recognition accuracy compared to the baseline system using MFCCs only. The improvement is observed at all noise levels for all noise types. Note the baseline here is the result from the Model-3mix, which showed the best performance amongst the models using MFCC+TV as shown in Fig 12. Also in Fig. 12 we show the performance of the 14 flesh-point pellet. The recognition accuracy here is averaged across all the noise types.

---

Frankel et al. [44, 45] noticed a significant improvement in recognition accuracy when the estimated articulatory data was used in conjunction with the cepstral features, which we also observed in our prior work [57]. We used the baseline MFCC features along with the estimated TVs for the ASR experiments. Here we considered three different models by varying the number of word (digit) mixture components per state from 2 to 4, identified as "Model-2mix", "Model-3mix" and "Model-4mix", where "Model-3mix" is the baseline model distributed with Aurora-2. Fig. 12 compares the recognition accuracy of MFCC+TVs from the different word models to the baseline accuracy using MFCC only. Adding TVs to MFCCs resulted in significant improvement in the word recognition accuracy compared to the baseline system using MFCCs only. The improvement is observed at all noise levels for all noise types. Note the baseline here is the result from the Model-3mix, which showed the best performance amongst the models using MFCC+TV as shown in Fig 12. Also in Fig. 12 we show the performance of the 14 flesh-point pellet. The recognition accuracy here is averaged across all the noise types.

---

We used this model for the rest of this paper.
3. Coding in ventral sensorimotor cortex (vSMC)

- Electrocorticography (ECoG) application of a mesh of tiny electrodes directly on the surface of the brain of a patient who is being prepared for brain surgery.
- Allows recording from very small populations of neurons.
- Developed at UC San Francisco in Edward Chang’s lab
- Examine multiple sites in vSMC while patient is speaking.
- Test which descriptions of speech best predict patterns of activation in particular electrode locations.

Encoding of electrode activity

- Possible predictors
  - Phonemes
  - Formant frequencies
  - Coordinated patterns of multiple articulators that form constrictions
  - Kinematics of individual articulators

- Because only acoustics are recorded, in order to evaluate articulators and constrictions, authors trained a model to infer time functions of EMA markers on lips, tongue, and jaw audio (overall correlation of original and inferred EMA = ~.65 for untrained speaker).

- Best predictor of electrode activity was EMA pattern associated with a gesture: coordinated articulator activity that produces and releases a constriction.

- Results consistent with Mugler et al., (2018) that show preference for gesture encoding over phoneme encoding in this area, and in IFG (inferior temporal gyrus, part of Broca’s area).
Weight pattern corresponds to coordinated articulator motion that produces and releases a coronal constriction.
Sites code distinct constricting effectors
What have we learned about constrictions as tasks?

- Evidence supporting constrictions
  1. Constrictions (Location and Degree) are specified by formant frequencies and amplitudes (Iskarous, 2010)
  2. Constriction information contributes to automatic speech recognition (Mitra et al., 2011)

- Evidence supporting other task variables
  1. Role of auditory information in speech production (Houde & Jordan, 1998)
  2. Evidence for lowered F3 as goal of /r/ production (Nieto et al. 2005)
  3. Patterns of neural activation in the motor cortex during listening code auditory patterns, not constrictions. (Cheung et al., 2016).
I. Role of auditory information in speech production

- Two sources of evidence for the real-time use of auditory feedback during speech production

- Auditory properties should be part of the task space (and forward map) of gestures

  **Sensorimotor adaptation** (e.g., Houde & Jordan, 1998)
  Sustained alteration of feedback can result in adaptation of speech production that is maintained after the alteration ends. (retuning of forward map)

  **Speech-induced suppression** (e.g., Chang et al., 2013)
  Response from some areas in auditory cortex (posterior superior temporal cortex) is reduced while speaking, compared to just listening to the same audio signal. (effference copy)
2. Evidence for lowered F3 as goal of /r/ production

- Articulatory variability in /r/ production is inversely related to the the effect of the variability on F3. (Nieto et al, 2005)
3. Patterns of neural coding in motor cortex during listening

- Activation in motor cortex during listening: evidence for constriction activation during speech perception?

- Selectivity of vSMC electrodes during speaking vs. listening (Cheung et al., 2016)

- Electrodes found that respond differentially to /b,d,g/ are typically those that are active only during speaking.

- Electrodes that are active during both listening and speaking do not exhibit clear selectivity as a function of constriction (labial, coronal, dorsal).

![Figure 2](image-url)

The following figure supplement is available for figure 2:
Figure supplement 1. Syllable token set.

DOI: 10.7554/eLife.12577
Clustering of electrode activation patterns

- vSMC
- Clustering by constriction type is strong during speaking.
- Clustering by constriction is weak during listening.
- Clustering by acoustic properties during listening is strong, but a bit weaker than in STG
- Clustering by Manner is stronger than clustering by constriction (place) during listening in vSMC.
- Even in the motor cortex, coding is auditory during listening.
Task space: going forward

- Formant frequencies and constrictions variables should both be part of task space

**Also** aerodynamic goals for acoustic sources:
- Pressures, laryngeal tension (McGowan & Saltzman, 1995)
- Voicing
- Turbulence?
- Intensity (for prominence)

- Since these goals (and formant frequencies) may result from action of more than one gesture, there needs to be a hierarchical organization of gestures into larger more global units to define such goals.

- Segments (constellations of gestures)? Problem is that the gestures composing the segment are often not synchronous.

- **When** do the global goals actively control the articulators ??

- Similar to problem arises with Constriction Degree hierarchy (Browman & Goldstein, 1989) as a model of sonority (propagation of CD from local constrictions to vocal tract as a whole, through parallel and serial concatenation of tubes). But when?
Related problem going forward: 
Articulatory-Acoustics Binding Problem

• Patterns of neural activation in motor cortex during listening and speaking are distinct.

• Constriction gesture responses during speaking

• Acoustics / manner responses during listening.

• Given evidence for sensorimotor integration (e.g., adaptive responses to altered auditory feedback) what binds them together?

• One possibility is change over time.

• The way the acoustics, auditory patterns, articulatory patterns change over time ought to be linked to one another.
Temporal Modulation functions

- Articulatory: Calculate the sum of the squared change in all the X-ray microbeam markers at every frame (kinetic energy of articulators)

- Acoustic: Calculate the sum of the squared change in all the MFCC at every frame (kinetic energy of spectrum)

- 23 speakers from XRMB corpus producing one sentence from ‘The Hunter’

- Modulation pulses
  - Modulation functions exhibit a repetitive sequence of pulses (7-9 Hz).
  - Generally one pulse per CV syllable (peak at C-V margin in spectrogram).
  - Additional pulse if there is a coda.

- Articulatory and acoustic functions are robustly correlated.
  - Better in narrow windows
  - Solution to binding problem?

\[
MFCC(k) = \sum_{i=1}^{13} (f(i, k + 1) - f(i, k))^2
\]

\[
MBEAM(k) = \sum_{i=1}^{7} \sum_{j=1}^{2} (m(i, j, k + 1) - m(i, j, k))^2
\]
Modulation pulses: some speculations

- Modulation pulses could provide temporal anchors for more global task goals: coordination of aerodynamic goals with respect to pulses.
- They could also provide a spatiotemporal basis for syllable structure.
- Highly speculative hypothesis: gestures are coordinated in such a way as to optimize a 7-9 Hz alternation in kinetic energy, potentially entrained to neural oscillations (theta?)
  - Could account for synchronization of onset C and V: Synchronization produces single maximum of kinetic energy, followed by a minimum.
  - Syllables obeying sonority sequencing should have clear modulation peaks separated by a minimum.
  - Could account for preferred sonority clines: Larger differences in sonority will result in higher kinetic energy peaks.
- Relation to rhythmic / mosaic structure
Prosody

- **Π gestures** (Byrd & Saltzman, 2003)
  - Model boundaries with a gesture whose task is to slow the speech production clock.
- Consequences of Π gesture application
  - Gestures are lengthened and slowed (reduced stiffness)
  - Reduced overlap of neighboring gestures
  - Lengthening of Vs and Cs at boundary
- Oscillators at different levels of the prosodic hierarchy that can be coupled (Tilsen, 2009)
  - Stress-timing vs syllable timing (O’Dell Nieminen, 1999)
  - Polysyllabic shortening (Saltzman et al., 2008)