USC Viterbi School of Engineering
USC University of Southern California
Ming Hsieh Department of Electrical Engineering

Movement Primitives in Speech Production

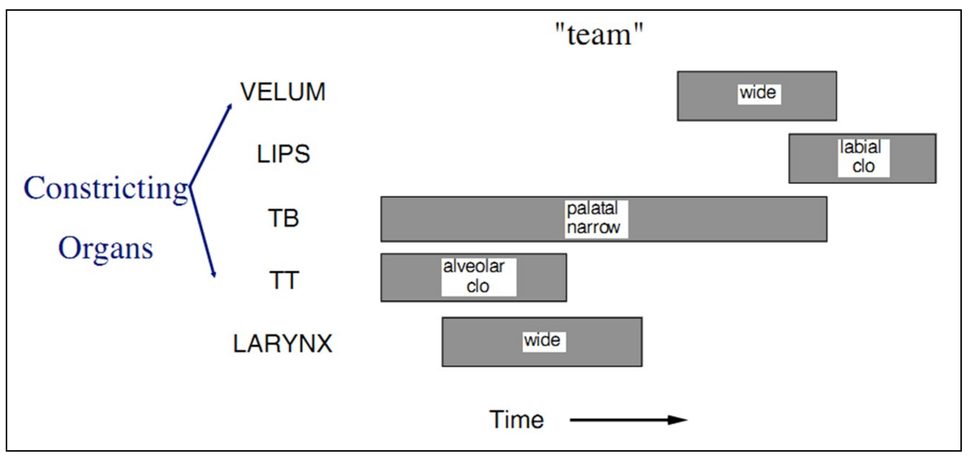
Vikram Ramanarayanan, EE-Systems

MOVEMENT PRIMITIVES: What are they & why do we care?

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.

- Aid explanation of variant and invariant aspects of articulation.
- Aid understanding of speech planning and execution at a cognitive level.

Theoretical ideas (Linguistics)

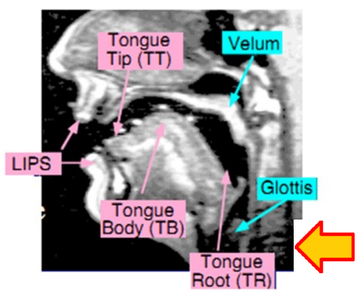


Gestural hypothesis:
Act of speaking can be decomposed into atomic units of action, or gestures.

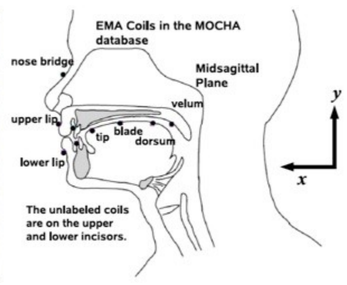
Gestures are dynamically-controlled constriction actions of distinct vocal tract organs. (e.g., lips, tongue tip, tongue body, velum, glottis)

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

Practice (Engineering)

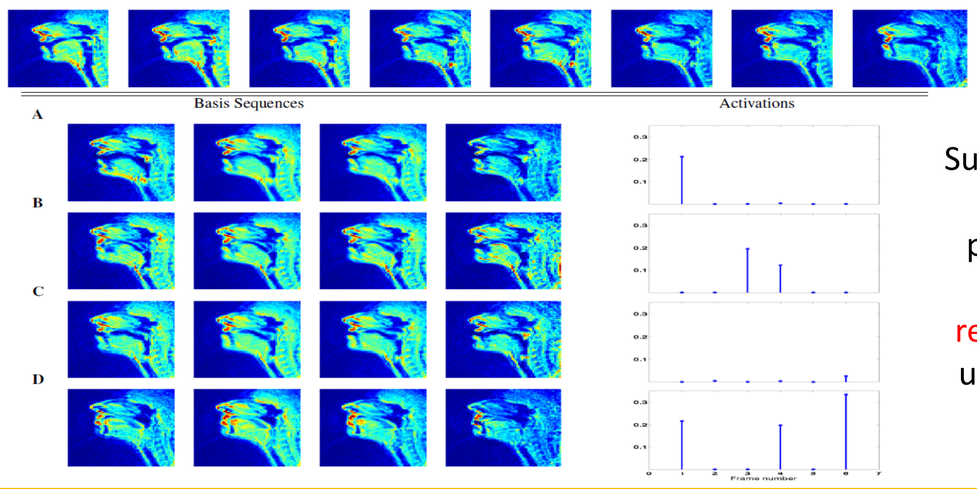


Electromagnetic articulography



Real-time magnetic resonance imaging

http://cvsp.cs.ntua.gr/research/inversion/



Such automatically extracted primitives allow for a compact representation of upper vocal tract motor actions

Extraction

Algorithm to *automatically* extract interpretable dynamic *movement primitives* from human speech production data .

1

$$\mathbf{V} = [I_1 | I_2 | \dots | I_N] \in \mathbb{R}^{M \times N}$$

Reformulate image sequence into a matrix for processing

2

Spatiotemporal basis M x K x T — Activation matrix K x N

$$\mathbf{V} \approx \sum_{t=0}^{T-1} \mathbf{W}(t) \cdot \vec{\mathbf{H}}^{t} = \mathcal{V}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad \overset{0\rightarrow}{\mathbf{X}} = \overset{\leftarrow 0}{\mathbf{X}} = \mathbf{X}$$

$$\overset{1\rightarrow}{\mathbf{X}} = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 4 & 5 \end{bmatrix} \quad \overset{\leftarrow 1}{\mathbf{X}} = \begin{bmatrix} 2 & 3 & 0 \\ 5 & 6 & 0 \end{bmatrix}$$

Convolutive NMF (nonnegative matrix factorization) is an elegant extension to ordinary NMF to obtain dynamic (time-varying) basis functions!

P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, no. 1, pp. 1–12, 2007.

3

Formulate the objective function subject to sparsity constraints as we don't want all bases to be activated at once:

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{V} - \sum_{t=0}^{T-1} \mathbf{W}(t) \cdot \vec{\mathbf{H}}^{t}\|^2 \text{ s.t. } sparseness(h_i) = S_h, \forall i.$$

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

$$sparseness(\mathbf{x}) = \frac{\sqrt{n} - \frac{(\sum_i |x_i|)}{\sqrt{\sum_i x_i^2}}}{\sqrt{n} - 1}$$

P. Hoyer, "Non-negative matrix factorization with sparseness constraints," The Journal of Machine Learning Research, 2004.

Interpretation & Validation

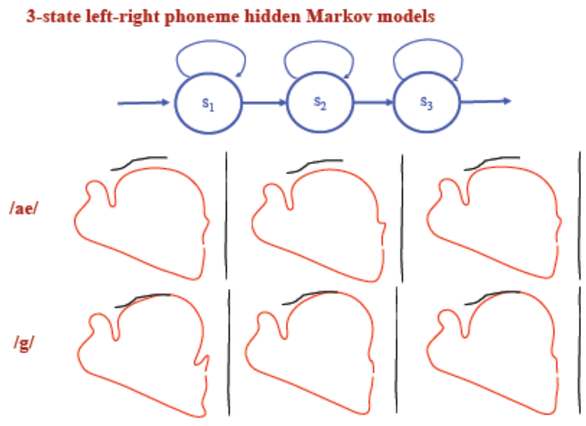
Validation is a difficult problem in general; no "ground truth" for real data!

1 Using synthetically-generated data:

- Use articulatory synthesizer (e.g.: TaDA) – generate synthetic data for which we can hypothesize "ground truth" bases and activations.

2 **Articulatory recognition experiments:**

- **3-state Hidden Markov Models (HMMs) built for broad linguistic classes (constriction degree and location) using the dynamic activation traces (columns of H over time) as features.**
- **Problem: HMM recognition not optimized for sparse features!**



Katsamanis et al., "Validating rt-MRI-based represent-ations using articulatory recognition," Interspeech 2011.

3 Significance of obtained H matrix:

- Do the obtained activation matrices afford a much smaller value objective function than any randomly generated matrix of the same sparsity?

Open Questions & Future Work

Algorithmic:

- Nonparametric and/or probabilistic extensions:
  - Don't have to fix parameters such as temporal dimension apriori
- Time-series clustering applications

Scientific understanding:

- Cognitive encoding of primitives
  - Links between production & perception
- Knowledge of primitives can inform optimal motor controller design

Work supported by NIH

Email: vramanar@usc.edu

Ming Hsieh Institute
Ming Hsieh Department of Electrical Engineering