statistical multi-stream modeling of real-time MRI articulatory speech data

Erik Bresch, Athanasios Katsamanis, Louis Goldstein and Shrikanth Narayanan
http://sail.usc.edu

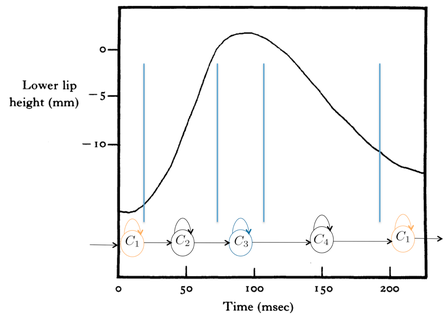
so, what?

Our goal is to computationally model speech production in a way that it would allow us to:

- gain deeper understanding of the underlying speech system by exploiting rich articulatory data (e.g., as acquired by rt-MRI)
- discover new or strengthen suggested links between articulatory observations and theoretical expectations

In this direction, we adopt the viewpoint of articulatory phonology [2] which is based on the description of an utterance "as an organized pattern of overlapping articulatory gestures".

lower lip gesture for the sound /aba/



A gesture is a goal directed action of constriction forming by a vocal tract articulator [2]

Our work focuses on:

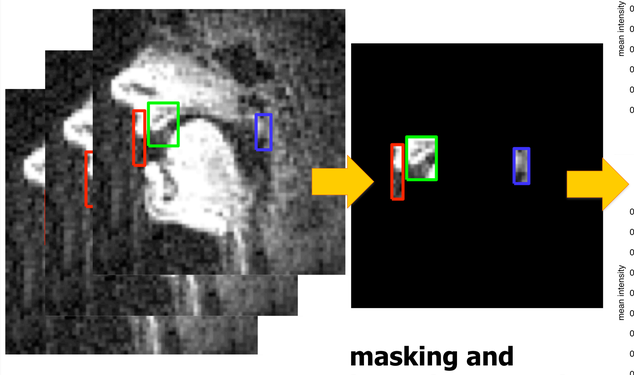
- extraction/measurement of articulatory gestures from rt-MRI sequences of the vocal tract
- multi-stream spatiotemporal articulatory modeling, that can account for gestural overlap and inter-articulator coupling

We model the couplings between three tract variables, i.e., lip aperture, tongue tip constriction degree and velum, using a coupled hidden Markov model.

Our model is verified in segmentation experiments of articulatory observations for previously unseen utterances. We compare with HMMs, GMMs, VQ.

The presented study indicates that if we combine a) an explicit multi-stream transcription (gestures) with b) appropriate techniques to extract articulatory trajectories form rt-MRI data and c) the appropriate statistical models, we are well-positioned to derive phonological information automatically from a rich set of articulatory data

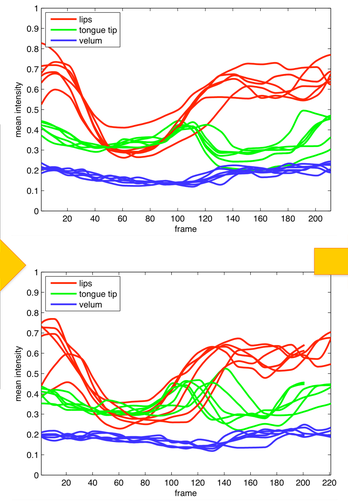
real-time Magnetic Resonance Imaging (rt-MRI) speech production data



lip, tongue tip and velum regions of interest

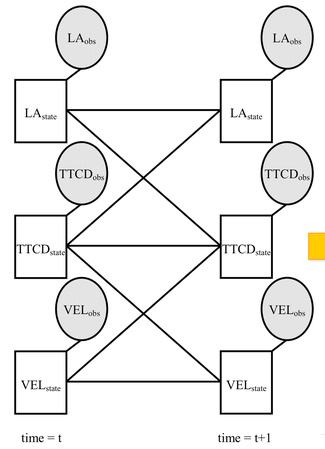
masking and intensity averaging

robust feature extraction



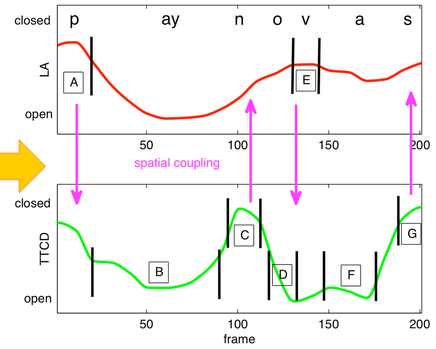
lip aperture (LA), tongue tip constriction degree (TTCD) and velum opening (VEL) traces

multi-stream, asynchronous, statistical articulatory modeling



coupled hidden Markov modeling

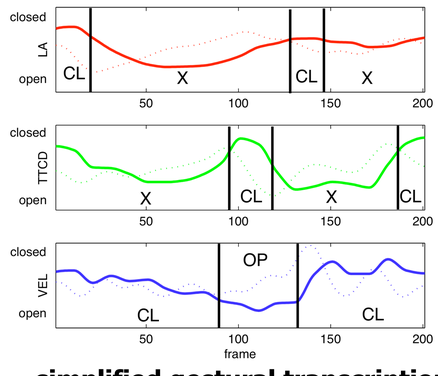
phonological information derivation and modeling



tract variable gestural event detection and segmentation

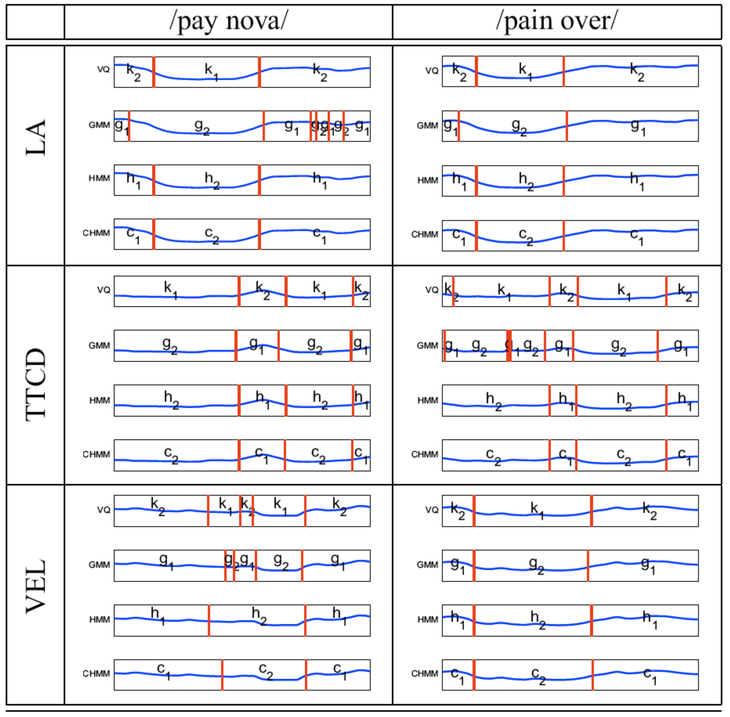
experiments

- ❑ **segmentation of tract variable gestures using Vector Quantization (VQ), Gaussian Mixture Models (GMM) , Hidden Markov Models (HMM) and Coupled Hidden Markov Models (CHMM)**
- ❑ **utterances (American English)**
  "type pay nova slowly" (7 realizations)
  "type pain over slowly" (7 realizations)
- ❑ **real time-MRI data at 22Hz**
- ❑ **leave-one-out cross validation**
- ❑ **verification: Nasal formation gesture synchronization in accordance with previous findings [12, 13]**



simplified gestural transcriptions

articulatory gestures' segmentation



references

[1] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," IEEE Signal Processing Magazine, May 2008.
[2] C. Browman and L. Goldstein, "Towards an articulatory phonology," Phonology Yearbook, vol. 3, pp. 219–252, 1986.
[3] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," Journal of the Acoustical Society of America, vol. 115, pp. 1771–1776, 2004.
[4] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," vol. 28, pp. 323–338, 2009.
[5] J. Fontecave and F. Berthommier, "A semi-automatic method for extracting vocal tract movements from x-ray films," Speech Communication, vol. 51, pp. 97–115, 2008.
[6] P. Badin, G. Bailly, L. Reveret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face based on MR and video images," Journal of Phonetics, vol. 30, pp. 533–553, 2002.
[10] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in Speech Production and Speech Modeling, W. Hardacastle and A. Marchal, Eds. Kluwer Academic Publishers,1990.
[12] D. Byrd, S. Tobin, E. Bresch, and S. Narayanan, "Timing effects of syllable structure and stress on nasals: a real-time MRI examination," Journal of Phonetics, vol. 37, pp. 97–110, 2009.
[13] E. Bresch, L. Goldstein, and S. Narayanan, "An analysis-by-synthesis approach to modeling real-time MRI articulatory data using the task dynamic application framework," 157th Meeting of the Acoustical Society of America, May 2009.

acknowledgments : This work was supported NIH Grant DC007124.