Project Summary
The vocal tract is the universal human instrument, played with great dexterity and skill in the production of spoken language. To produce the elegant acoustic structure of speech, the ongoing shaping of the vocal tract must be choreographed with moment-to-moment precision, as it is this vocal tract shaping, in coordination with respiratory and laryngeal behavior, that produces not only the linguistically significant resonant structure of vowels and sonorant consonants and the complex aerodynamic qualities of obstruent consonants but in endowing speech with the rich expressive emotional quality. Little has been systematically and quantitatively established, however, about the ways in which how individual vocal morphological differences are reflected in the acoustic speech signal and what articulatory strategies are adopted in the presence of structural differences to achieve speech invariance. The overarching objective of the proposed research is to improve our scientific understanding of how vocal tract morphology and speech articulation interplay and explain the variant and invariant aspects of speech signal properties within and across talkers. A related goal is to create forward and inverse models that relate vocal tract details to the resultant acoustics that can shed light on individual differences during speech.

The state of art in speech research has predominantly focused on surface speech acoustic properties; there remain open questions as to how speech properties co-vary across talkers, linguistic and paralinguistic conditions. Given the complex interplay between these factors, there are limitations to uncovering the underlying details just from the resultant acoustics. This proposal focuses on direct experimental and computational investigation of the dynamic human vocal tract using novel imaging techniques and signal modeling to illuminate individual structural vocal tract variability as well as the strategies by which linguistic articulation gets implemented. Using two vocal tract data acquisition systems—three-dimensional Electromagnetic Articulography that offers excellent temporal resolution and a real-time Magnetic Resonance Imaging with superior spatial resolution of the entire moving vocal tract that we helped develop (Fig 1)—we plan to gather and quantify the spatio-temporal details of speech production across individuals. The initial phase of the study will focus on data from 30 individuals.

We adopt the gestural framework of Articulatory Phonology for our modeling due to its theoretical rationality of the “gesture” concept that offers a compositional view of speech production, its dynamic nature and, critically, the availability of an articulatory synthesis tool in that framework. The proposal aims to understand, as a function of vocal tract structure, (1) the ways speakers modulate the spatiotemporal organization of syllable articulation and (2) the ways such articulatory variations interact with speech acoustic properties. Two concurrent computational goals will be undertaken empowered by the novel direct vocal tract data. The analysis-by-synthesis approach to modeling with an articulatory synthesizer will allow for discovering realistic rules for emotional speech synthesis. To do this, we will employ the TAsk Dynamic Application (TADA) model of Haskins Laboratories, a computational implementation of Articulatory Phonology and Task Dynamics. The model is particularly appropriate because it provides a hypothesized ensemble of gestures arrayed over time for any input utterance. The model is biologically plausible and produces as its output explicit time-functions of constriction events in the vocal tract, which is precisely what we measure directly with real-time MRI. We will also investigate how well these articulatory features can be estimated from the speech signal using acoustic-articulatory inversion.

Investigating how individual specific information is encoded in speech is important from a variety of viewpoints both scientific (in advancing our understanding of the basic mechanisms of human speech generation) and technological (in creating computational tools for the analysis and recognition of speech).