The Gestural Structure of Speech

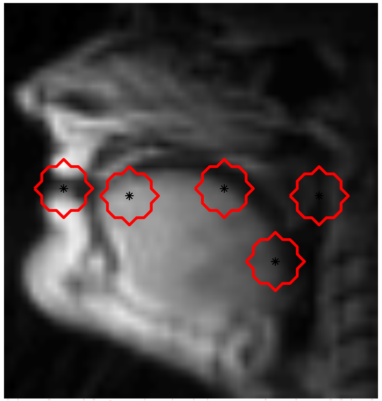
Phonetic Theory

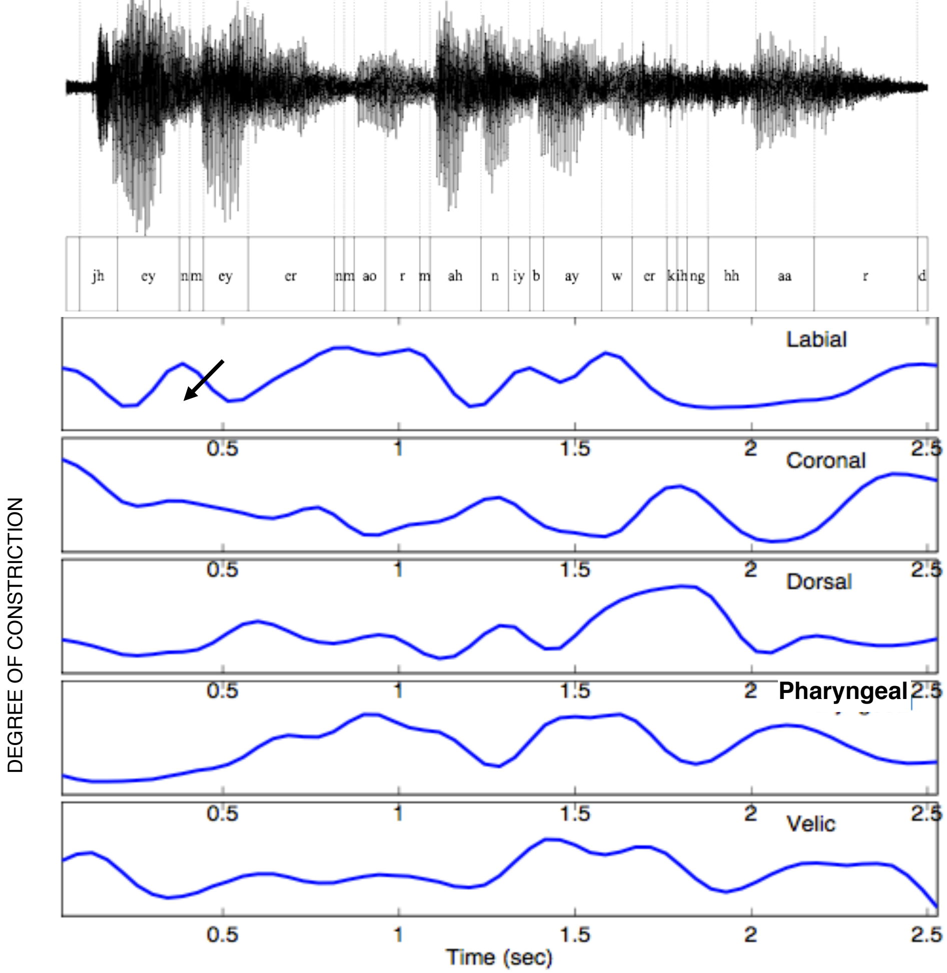
- Phonetics aims to develop:
  - a theory of the structure of the human speech system
  - how it is functions in communication
  - how it is cognitively represented
- Satisfying these aims require characterizing:
  - the continuous properties of speech articulation, acoustics, aerodynamics, audition, etc.
  - AND their communicative, informational and cognitive properties
- This leads to a classic theoretical problem.

The Problem:

Apparently incompatible descriptions of speech

“Jane may earn more money by working hard”





- Phonological (Informational/Cognitive)
  - 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

Phonological Units	Physical Measurements
Discrete	Continuous
Time-invariant	Time-varying
Context-independent	Context-dependent
Low Dimensionality	High Dimensionality

- Historically (e.g. Hockett, 1960) the incompatibility is "resolved" by separating cognitive and physical descriptions.
- But this just pushes back the problem.

Articulatory Phonology (Browman & Goldstein, 1989): Key components of resolution

Gestures

- Dynamical systems representation
- Articulatory Synergy representation
- Coordination and Overlap

Phonological Units	Physical Measurements
Discrete	Continuous
Time-invariant	Time-varying
Low Dimensionality	High Dimensionality
Context-independent	Context-dependent

Dance of the Vocal Organs

Information and the dance

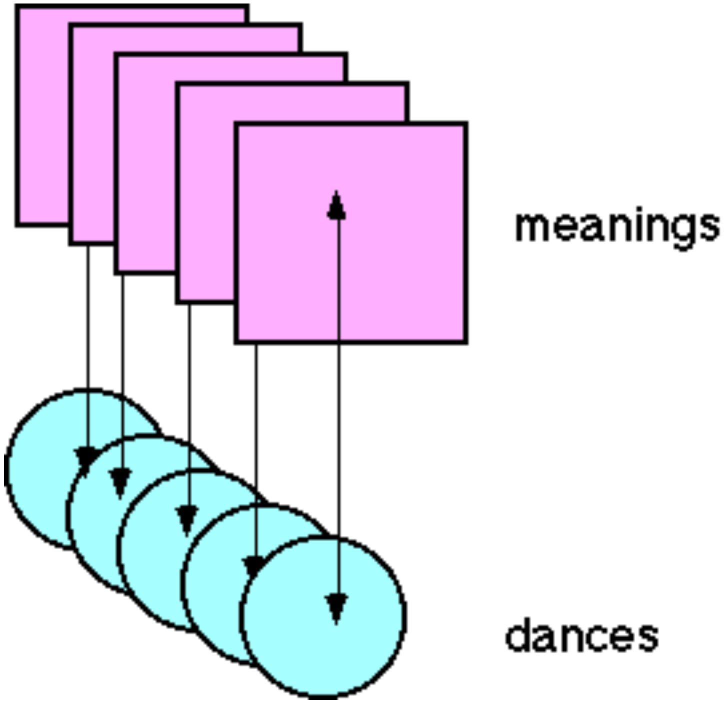
- All the information in our messages (thoughts, ideas, etc.) of arbitrary complexity must ultimately be associated with unique simple dances.
- We can convey a potentially infinite number of different messages… infinite number of dances.
- Systems in nature that make use of finite means to produce an infinite number of distinct messages all employ discrete units that can be arranged in an infinite number combinations (syntax, genetics, chemistry…)
- What are the discrete units of the dance??

Words and Contrast

- Words (or, morphemes) are elements of a language that have distinct meanings and which are also associated (arbitrarily) with distinct "dances" of the vocal tract organs.

- This is the informational or contrastive function of the dance.

- Despite differences among (and within) individual speakers, the discrete units of the dance must be shared for all speakers of a given language community.



- What are the discrete units?
- How do individuals perform them differently?

Gestures as *units of action*

- Hypothesis:
  Just as dance can be decomposed into discrete actions (steps) that are choreographed, the motion of the vocal organs in speech can be decomposed into discrete actions or gestures (that are also choreographed):

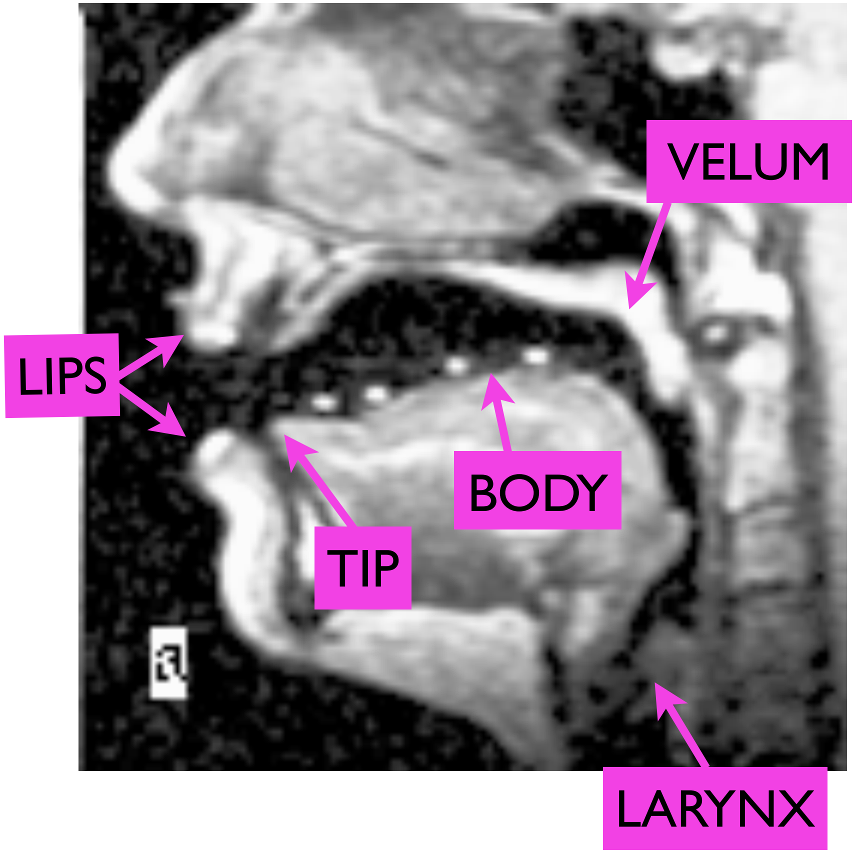
- A gesture is...
  - a constriction action of one of the distinct vocal tract organs.
  - For example, the words "bad", "pad", "meter", and "Bingo" all begin with closure gesture of the lips organ.
  - We observe that in producing such words, speakers' upper and lower lips always come together to form tight seal.
  - This closure action is an essential property of the dance for these words.

Gestures as *units of information*

- Gestures are themselves meaningless, but they function to distinguish words (minimal messages units) from one another.

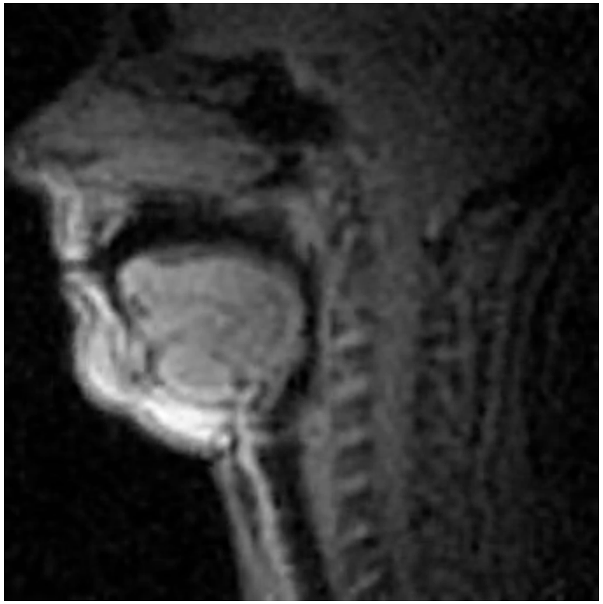
- Gestures of distinct constricting organs can be used to distinguish words from one another in all languages (intrinsic discreteness).

- For example:
  - "bad" begins closure gesture of the lips organ.
  - "dad" begins with a closure gesture of the tongue tip organ.
  - "gal" begins with a closure gesture of the tongue body organ.

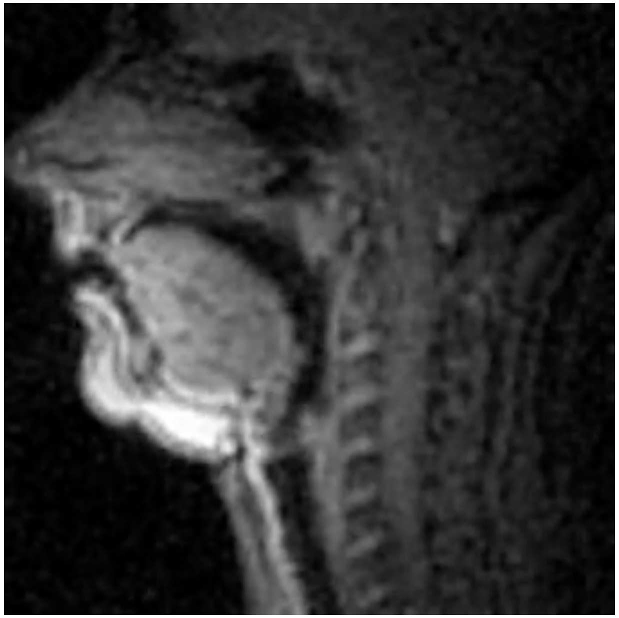


Oral Constriction Gestures

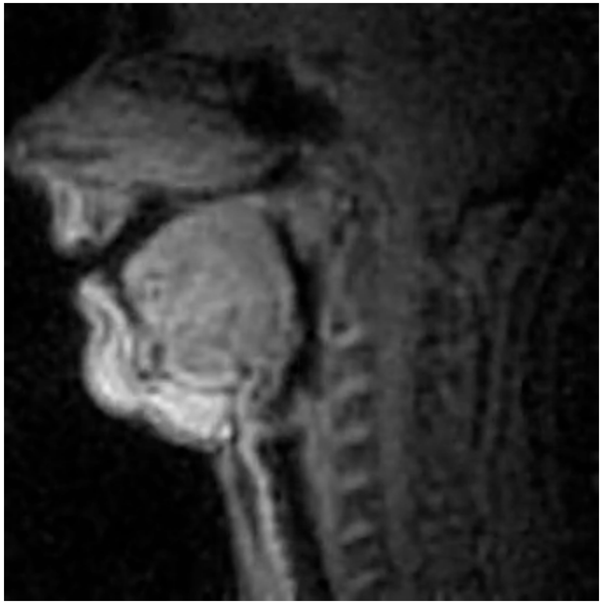
LIPS



Tongue Tip

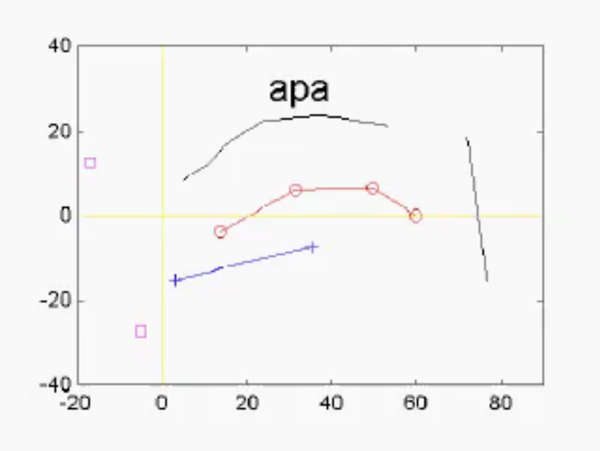


Tongue Body

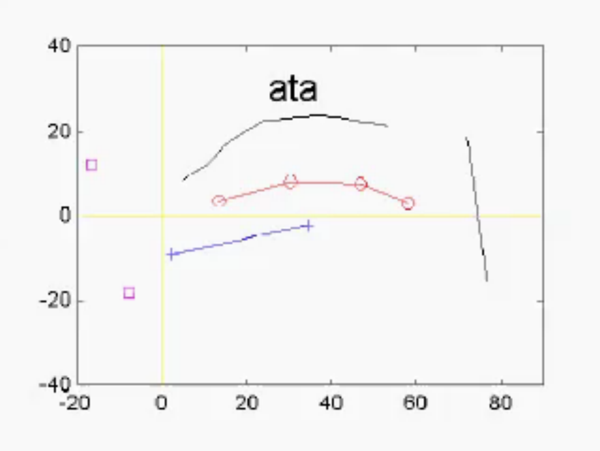


Constriction Gestures: markers

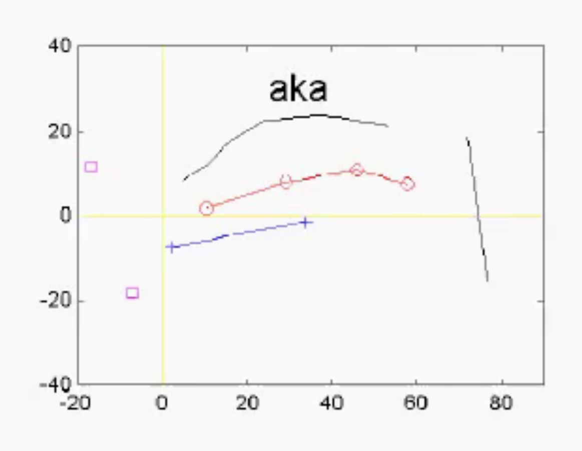
LIPS



Tongue Tip



Tongue Body



Tasks, articulators, redundancy, synergy

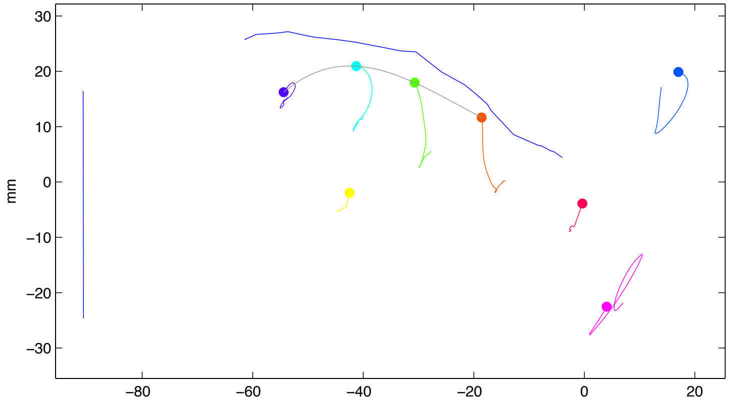
- Performance of any skilled motor task requires cooperation of several independently moveable body parts, which we call articulators.
    - e.g., reaching for an object on a table
- There is large (possibly infinite) set of articulator postures that will achieve the task. This is sometimes called redundancy.
- 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 a control law.
- The synergy allows the task to be performed in different ways in different environmental contexts.
- Different actors learn to "tune" the synergy differently, resulting in different articulator movements for the same functional task.

Harnessing Redundancy

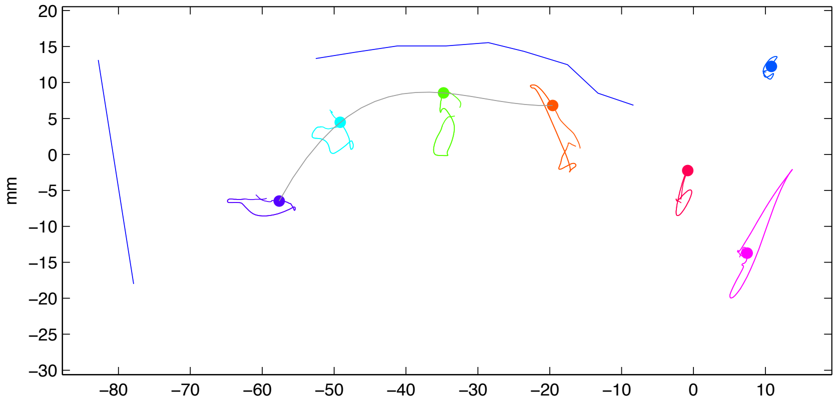
Synergies in Speech

- Tasks in speech are constrictions that form the consonants and vowels: the task(s) of a gesture are discrete, context-independent.
- e.g., task that is in common to /p,b,m/ is the closure of the lips. In other words, reduce the distance between the lips (Lip Aperture) to 0.
- What articulators are part of the synergy for a lip closure?
  - jaw
  - lower lip
  - upper lip
- Different people learn to tune the synergy differently: They employ different relative contributions of these articulators.
- Relative contributions differ when some perturbing events occur in the world.
- Relative contributions differ when the task is produced in the context of other tasks.

Speaker differences in Lip Closure synergies

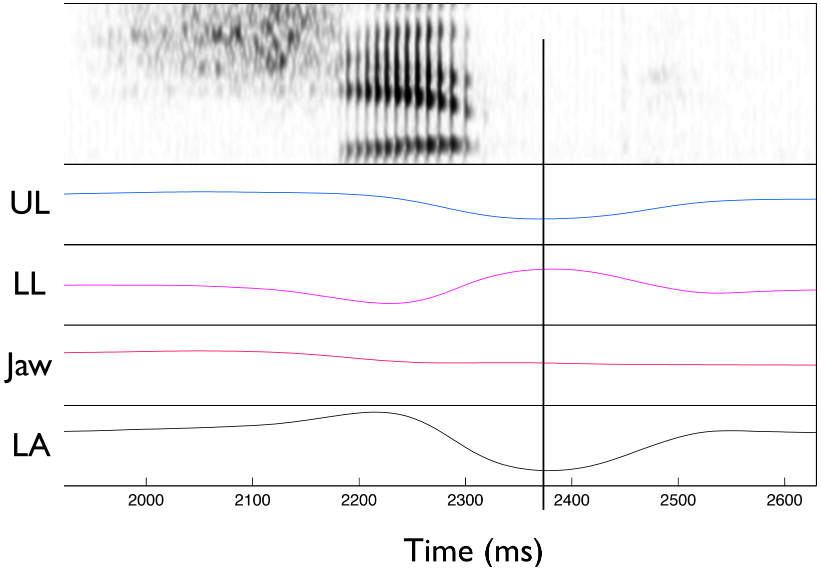


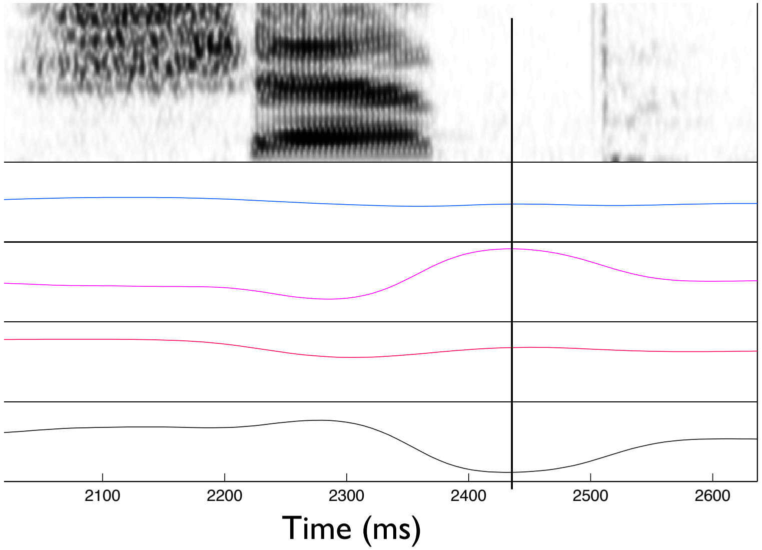
“ship’



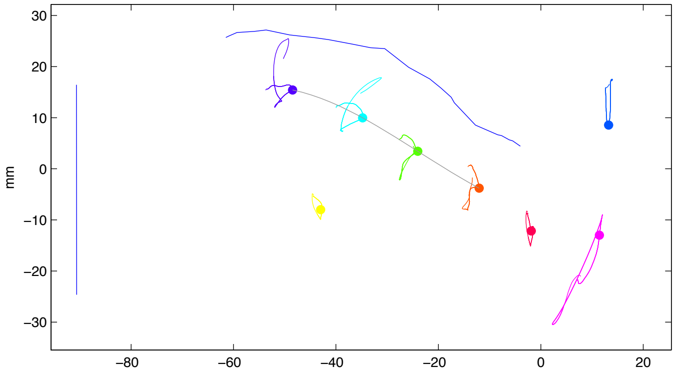
Speaker A

Speaker B

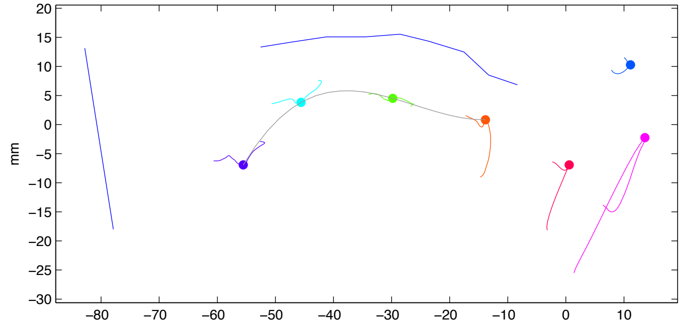




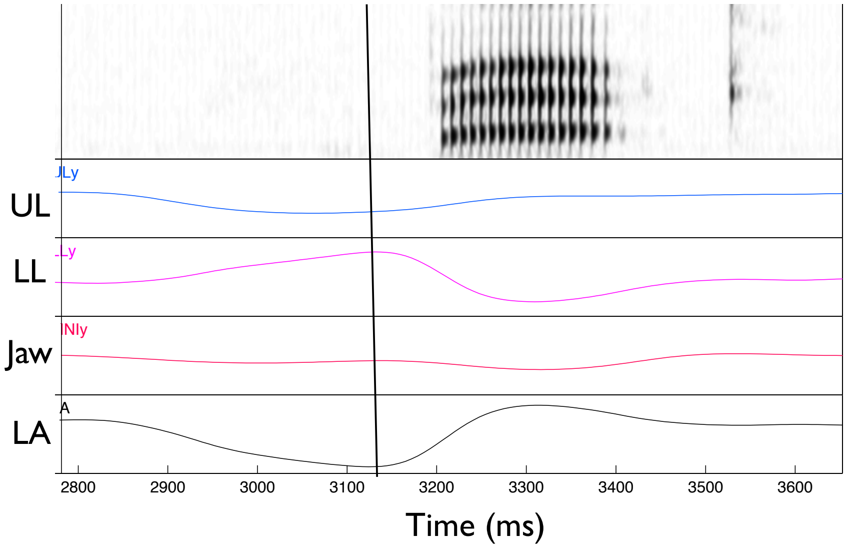
Speaker differences in Lip Closure synergies



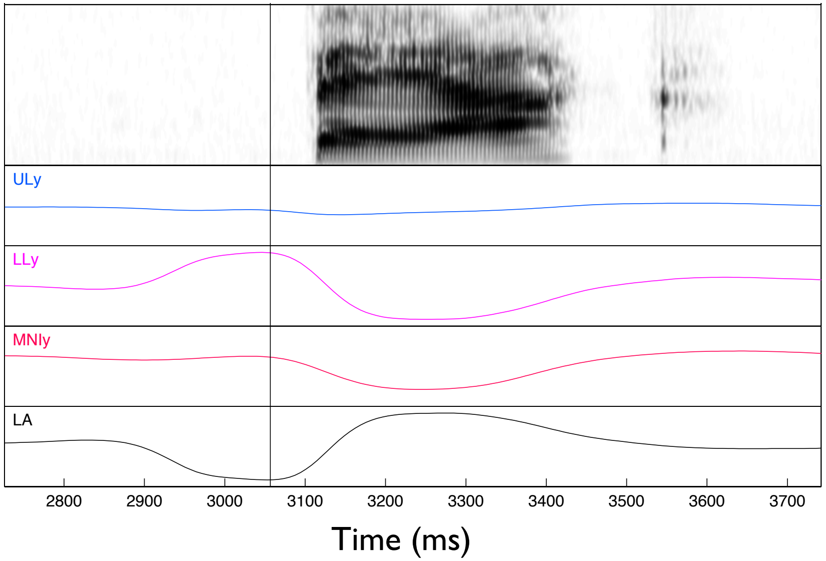
“back”



Speaker A



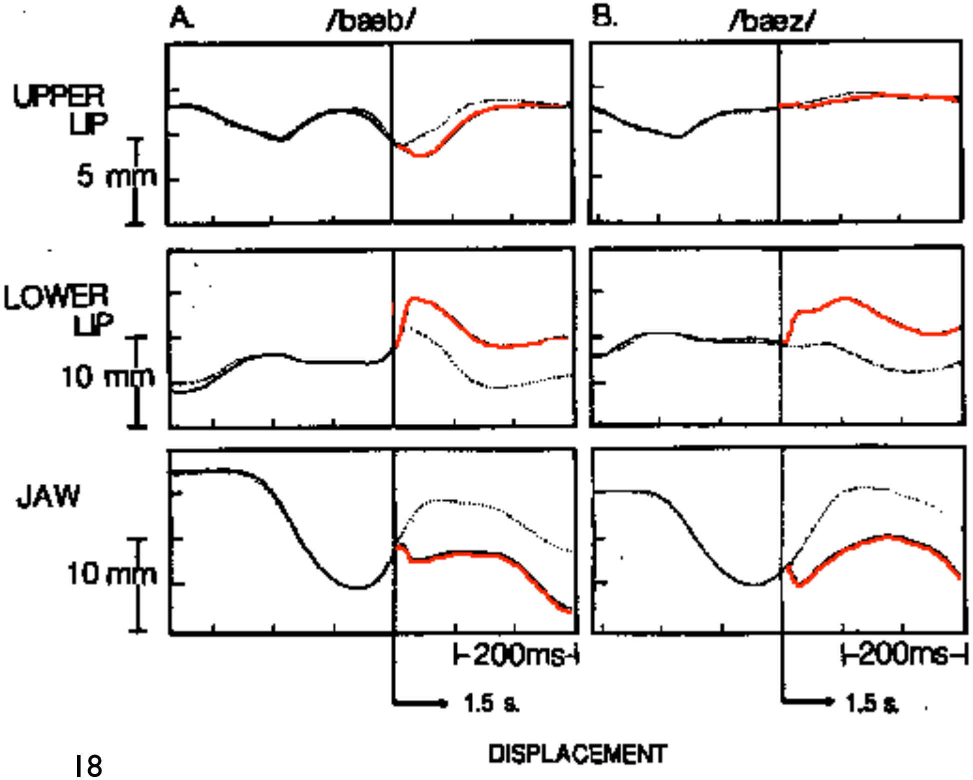
Speaker B



Compensation for perturbation

- Compensation
  When the task is threatened by a perturbation of one articulator (e.g., yanking on the speaker's jaw as (s)he is about to produce a lip closure), other articulators, remote from the site of the perturbation, act to to meet the challenge (e.g., increased displacement of upper lip) (Kelso et al, 1983).

- Speed
  Compensatory action is extremely fast (20 ms or so). This implicates direct inter-articulator cooperation. There is not enough time for an executive to "manage" responses to perturbation.
- Task-specificity
  Response to perturbation is task-specific, not hard-wired. If the subject is producing /z/, instead of /b/, response is not seen.



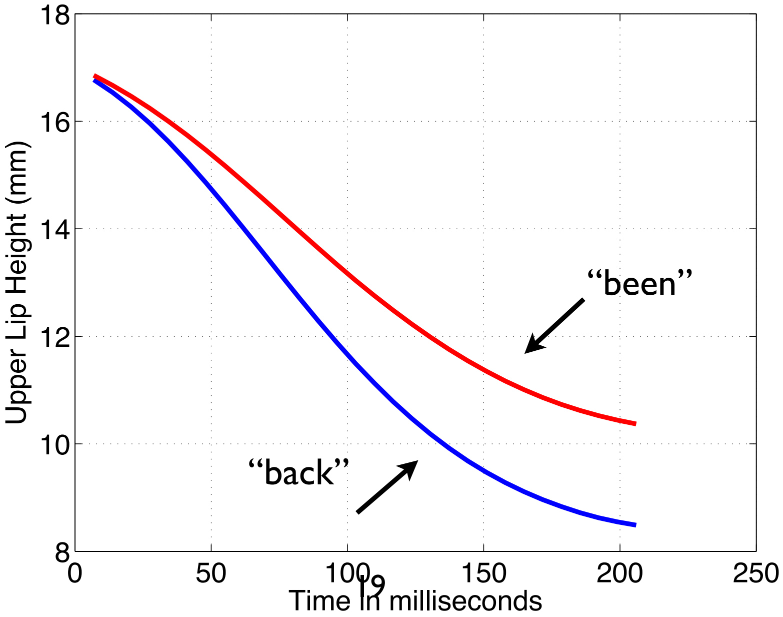
Lip Task performance in different contexts

- The relative contribution of the articulators in the synergy may differ when the task is produced in different contexts in which one of the articulators may be required for some other task.

- For example, lip closure in “back” vs. “been”.

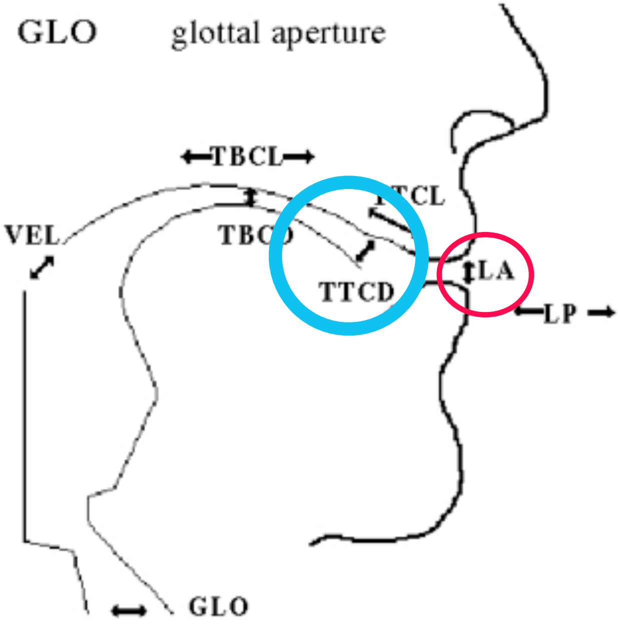
- Jaw is recruited to be low in “back” because of the low vowel (/ae/) and high in “been” because of high vowel (/ɪ/),

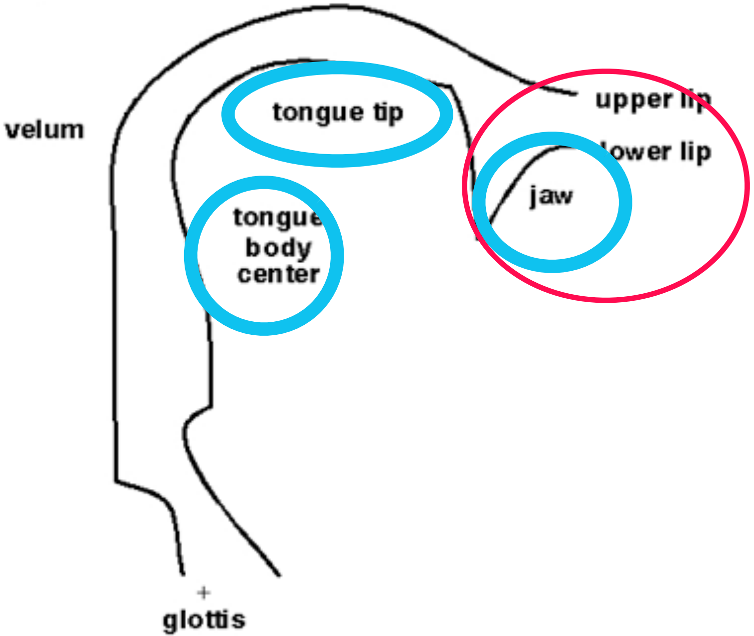
- More upper lip lowering emerges in “back” than in “been”.



Gesture task variables

Task		Articulators
LP	lip protrusion	upper & lower lips, jaw
LA	lip aperture	upper & lower lips, jaw
TTCL	tongue tip constrict location	tongue tip, tongue body, jaw
TTCD	tongue tip constrict degree	tongue tip, tongue body, jaw
TBCL	tongue body constrict location	tongue body, jaw
TBCD	tongue body constrict degree	tongue body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

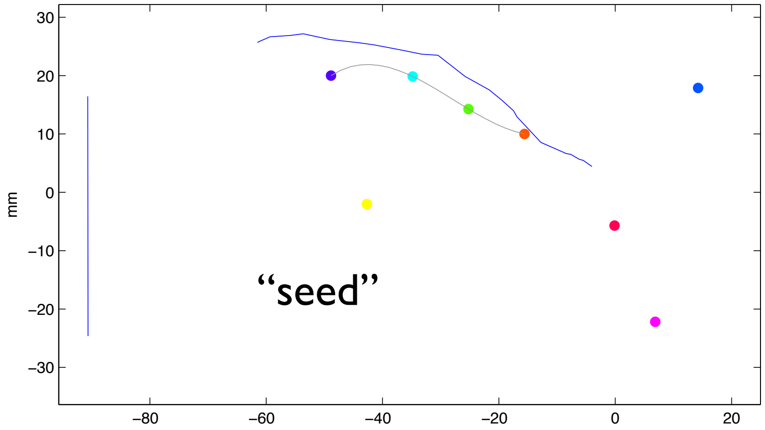


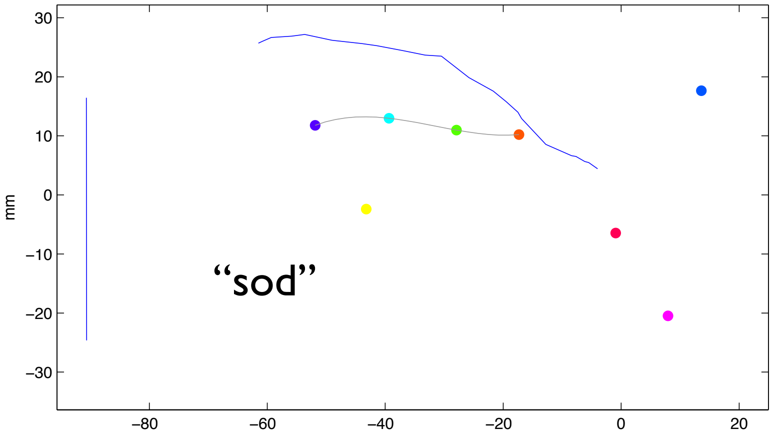


Tongue Tip Task performance in different contexts

- In the context of different vowels, the tongue tip closure task for (/t,d,n/) is produced with a different combination of articulators: tongue body, tongue tip.

- This is sometimes called "coarticulation"





Speech as audible gesture

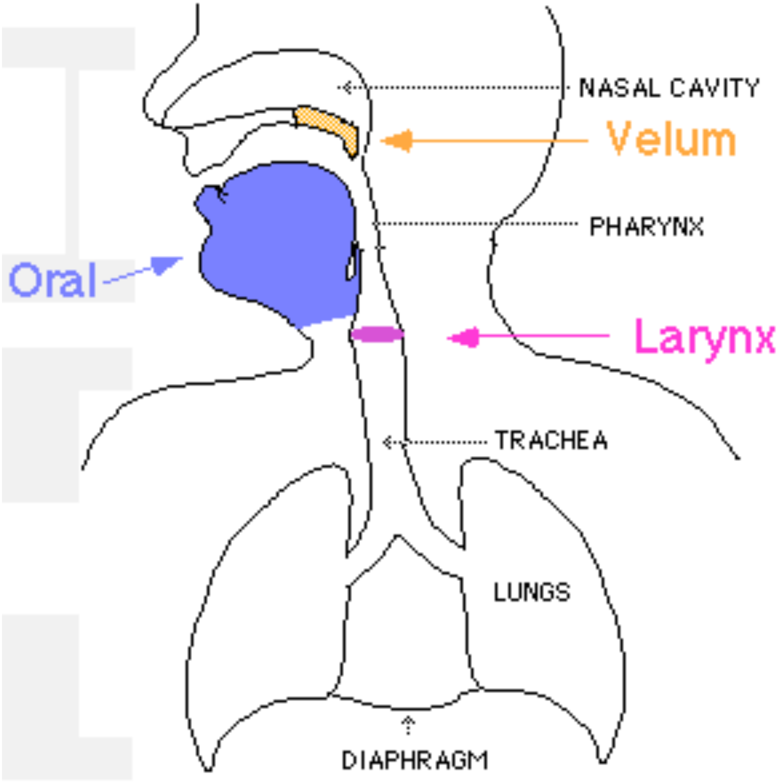
- Gestures are analogous to the use of "signs" in languages such as American Sign Language (ASL).
- In sign, gestures of the arms, hands, and fingers are communicated optically to the visual system of the receiver.
- In speech, gestures of the tongue, lips, and larynx are (largely) invisible, but are communicated acoustically to the auditory system of the receiver.
- For gestures that are potentially visible, optic and auditory information are combined into a single gestural percept.

Gestures and sound production

- Two functions of sound production need to be distinguished:
  - Sound generation
  - Sound shaping
- Sound generation: causing air to vibrate at audible frequencies
  - In the case of musical instruments, this is the function of lips against the mouthpiece in trumpet, or the air passing over the reed in a wind instrument.
  - Device generating sound is the sound source.

Sound sources in speech

- vibration of the larynx
- turbulent ("jet") noise of air rushing thru narrow slit
- "pop" when built-up pressure is released



Sound shaping

- Generated sound is shaped, or filtered, by passing it through tubes of various lengths.
- Tube lengths determines the spectrum of the sound, the relative strength of the frequencies or overtones that the tube vibrates at.
- In trumpets or wind instruments, different shaping is accomplished by fingering.
- In most mammals, the length of vocal tube is used as information about an animals size.
- Filter functions in speech:
  - constrictions of different organs produce changes in the effective lengths of vocal tract tubes.
  - allowing air to pass through the nose or not.

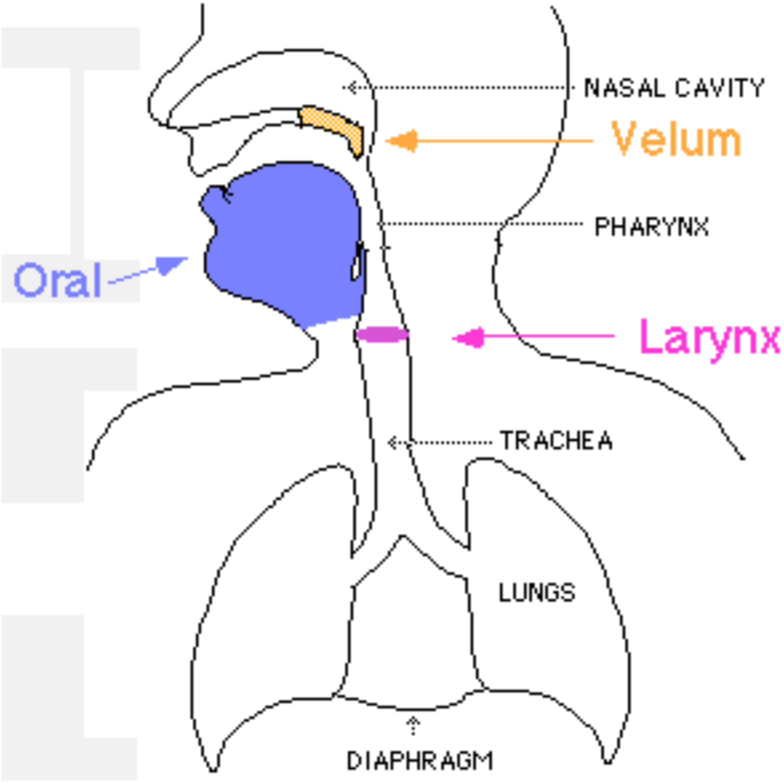
Speech Gestures and sound

- Gestures of functionally distinct constricting organs can distinguish words:

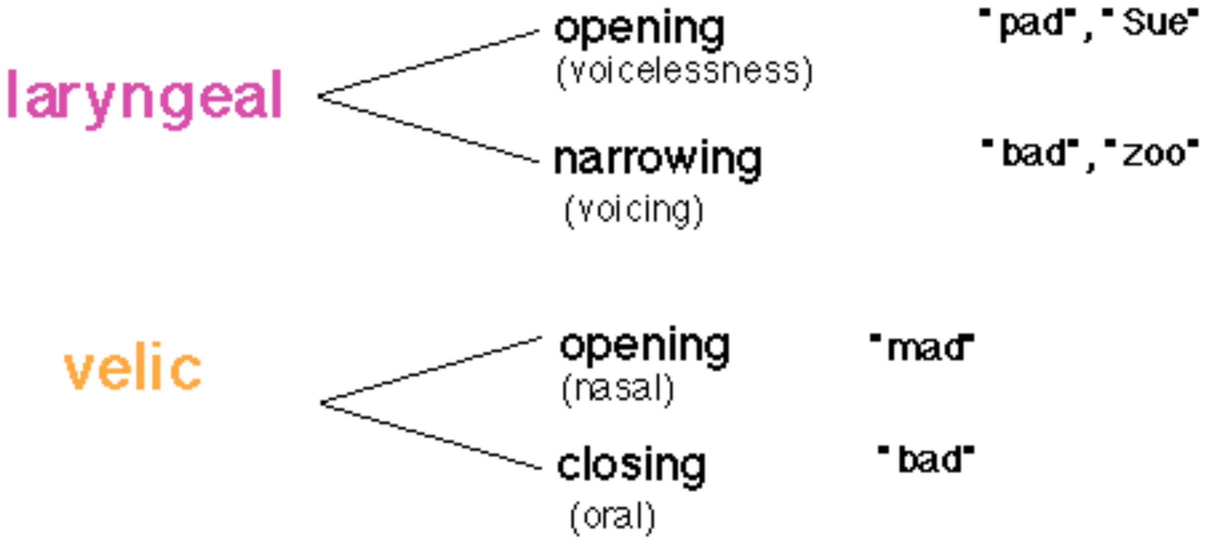
- Larynx (generates sound source)

- Velum (shapes sound generated at larynx)

- oral constrictors: (shapes sound generated at larynx)
  - lips
  - tongue tip
  - tongue body



Types of Gestures



laryngeal

opening (voicelessness) "pad", "Sue"

narrowing (voicing) "bad", "zoo"

velic

opening (nasal) "mad"

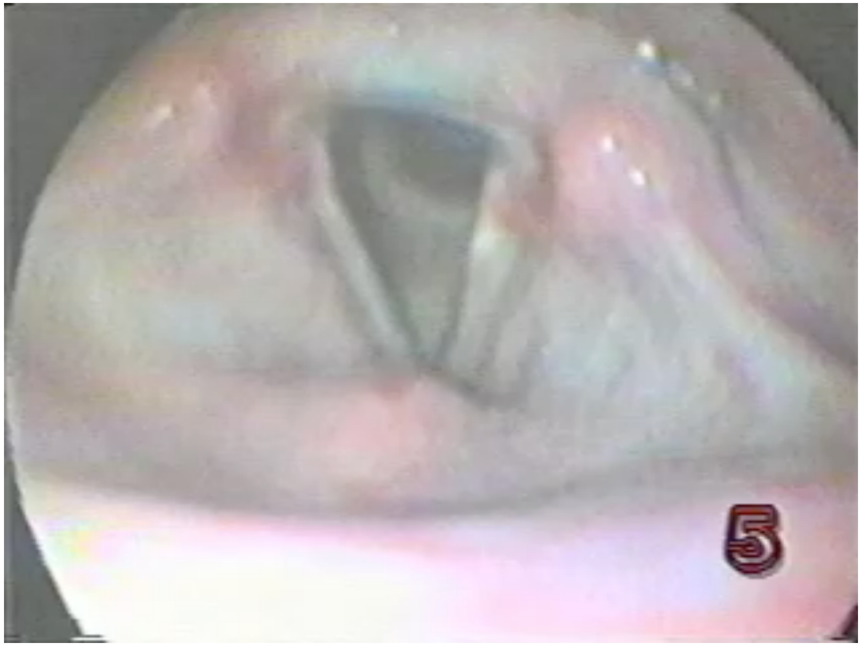
closing (oral) "bad"

oral

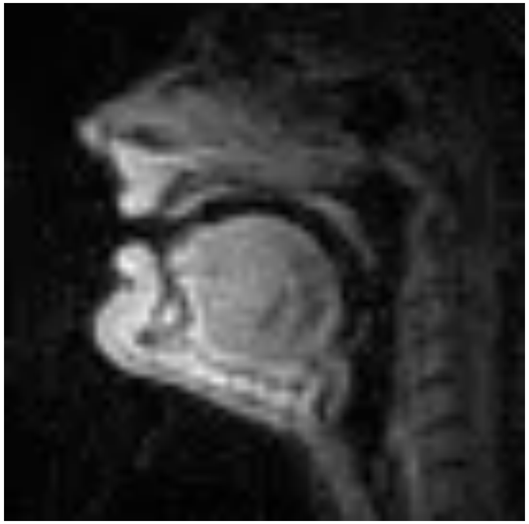
Constricting Organs:

Lips	Labial	"bought"
Tongue Tip	Coronal	"dot"
Tongue Body	Dorsal	"got"
Tongue Root	Radical	"rot"

Laryngeal Gestures

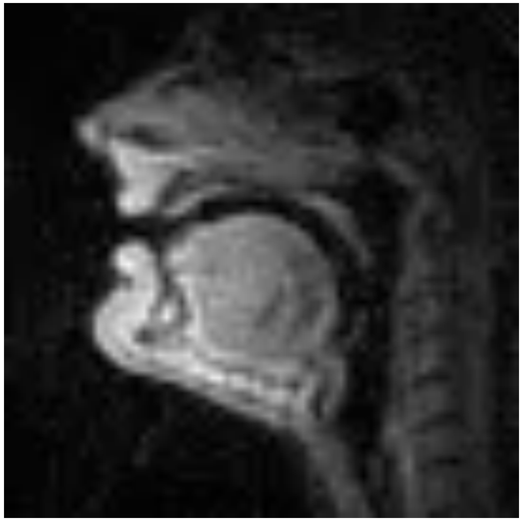


Velic Gestures

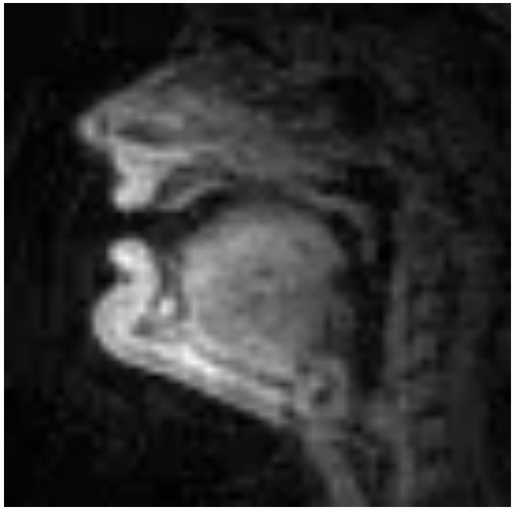


"Jane may earn more money by working hard."

Velic Open:
NASAL



"Jane"

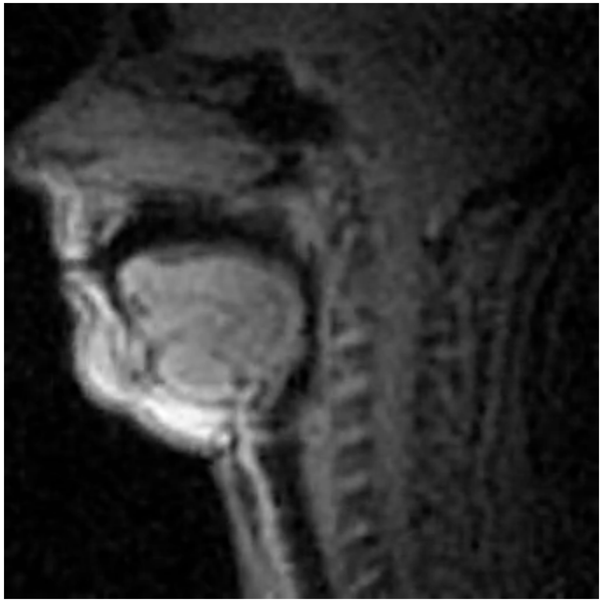


"hard"

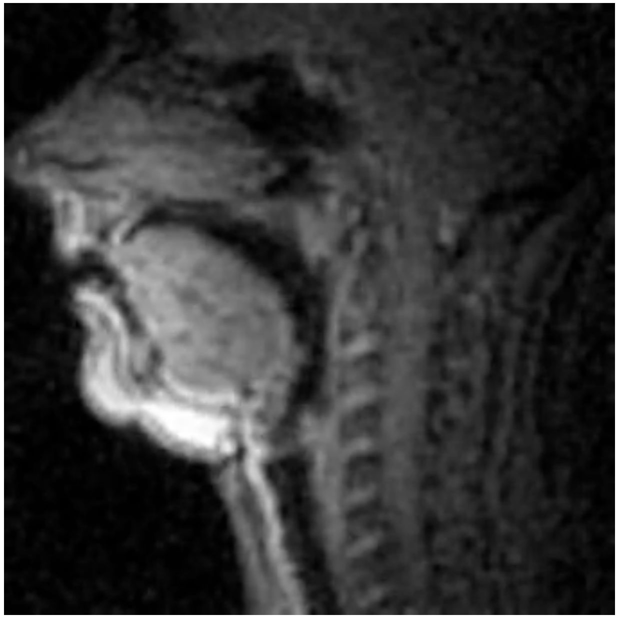
Velic Closed:
ORAL

Distinct Oral Constriction Gestures

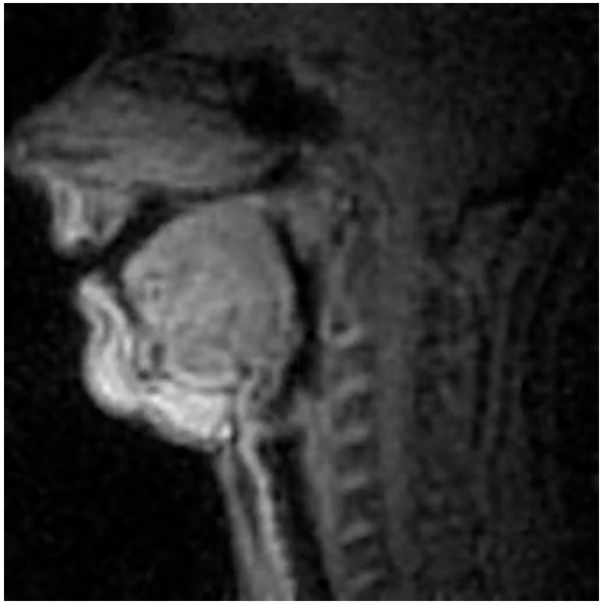
LIPS
Labial



Tongue Tip
Coronal



Tongue Body
Dorsal



These contrast in every language

Gesture Combinations

- English words can begin with combinations of Oral, Laryngeal and Velic constriction gestures.
- The resulting combinations are usually analyzed as consonants or consonant segments.
- From the gestures we illustrated, we can form 9 combinations (consonant segments) in English.

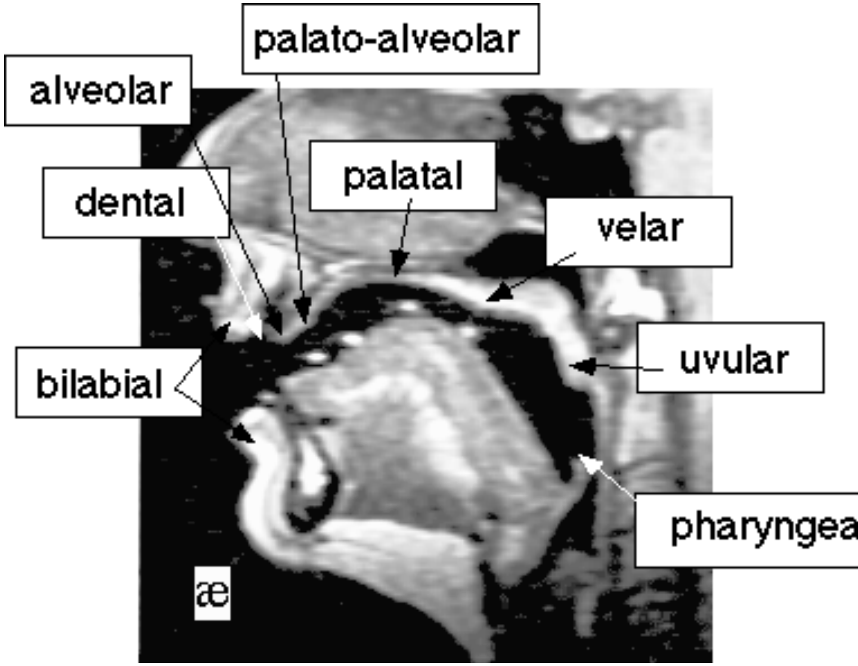
VELIC	closed	closed	open
LARYNX	narrow	open	narrow
LIPS	“bought”	“pot”	“Mott”
TT	“dot”	“tot”	“not”
TB	“got”	“cot”	“pong”

- But there are more than 9 consonants in English. Where do the rest come from?

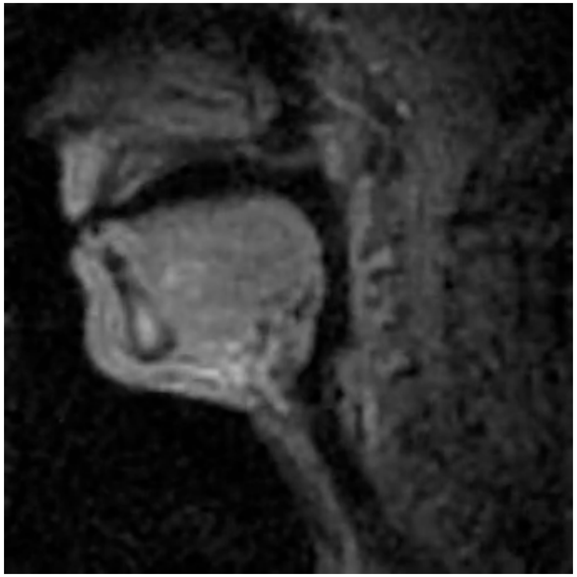
Differentiating oral constriction gestures

A given constrictor can produce several different contrasting gestures by varying the parameters of the task goals. (0 is not the only goal)

- Constriction Degree
  (e.g. goal of TTCD: how narrow is the constriction?)
  - stop ("dip, tip")
    complete obstruction of tube
    generates "pop" sound source
  - fricative ("zip, sip")
    narrowing to create jet noise source
  - approximant ("rip")
    narrowing with no source change
- Constriction Locations
  (e.g., goal of TTCL: exactly where is is the constriction formed?)

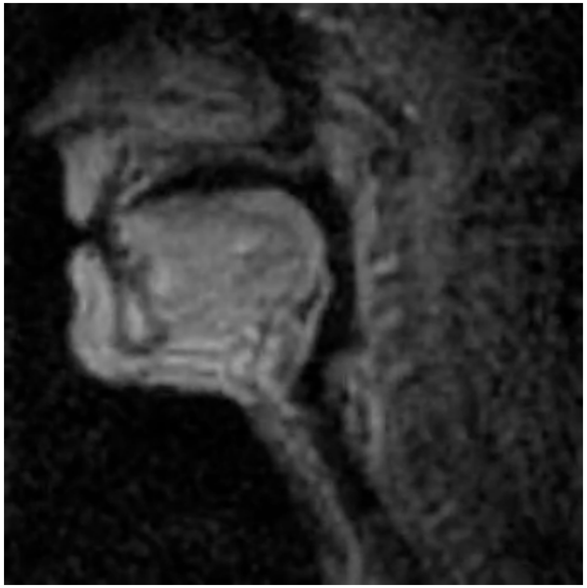


Constriction Locations for TT fricatives



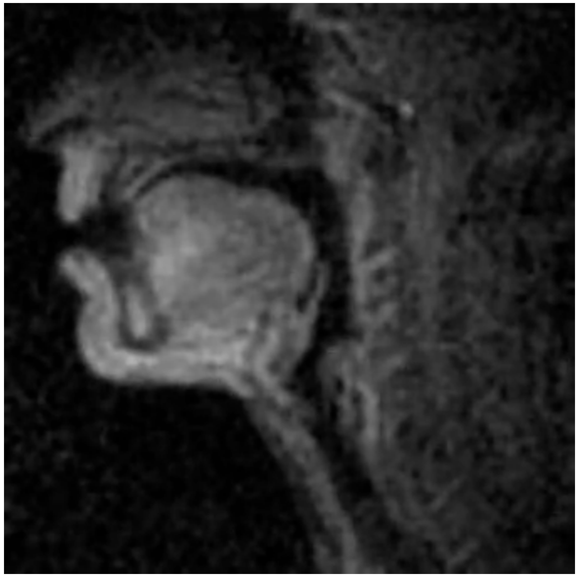
dental

[aθa]



alveolar

[asa]



palatoalveolar

[aʃa]

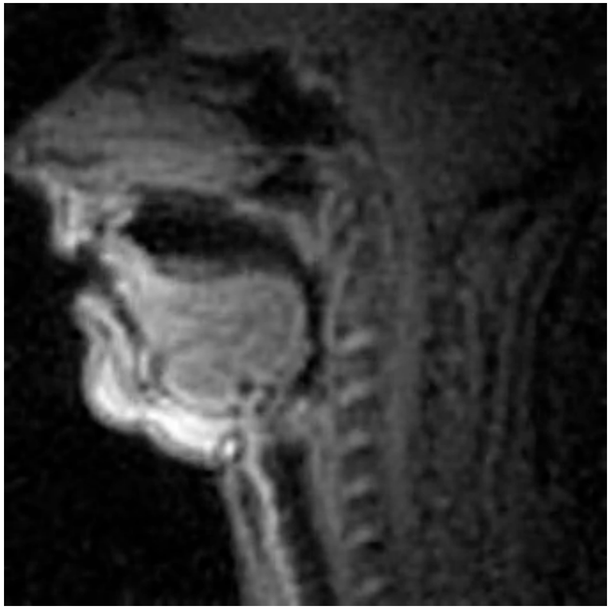
Gestures vs. Features

- Gesture tasks seem similar to features.
- Considering only the static task description that is true.
- But there is a critical difference:
  Gestures are events that unfold over time lawfully,
  according to their dynamics.
- Features do not solve the incompatibility problem.

Multiple oral constrictions

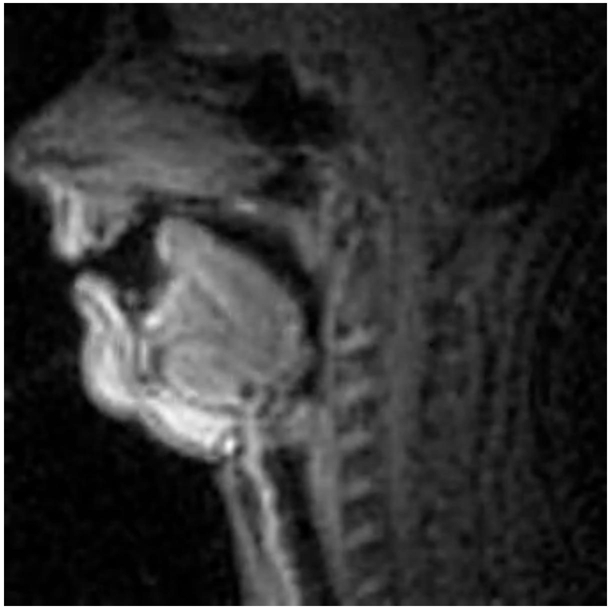
“lie”

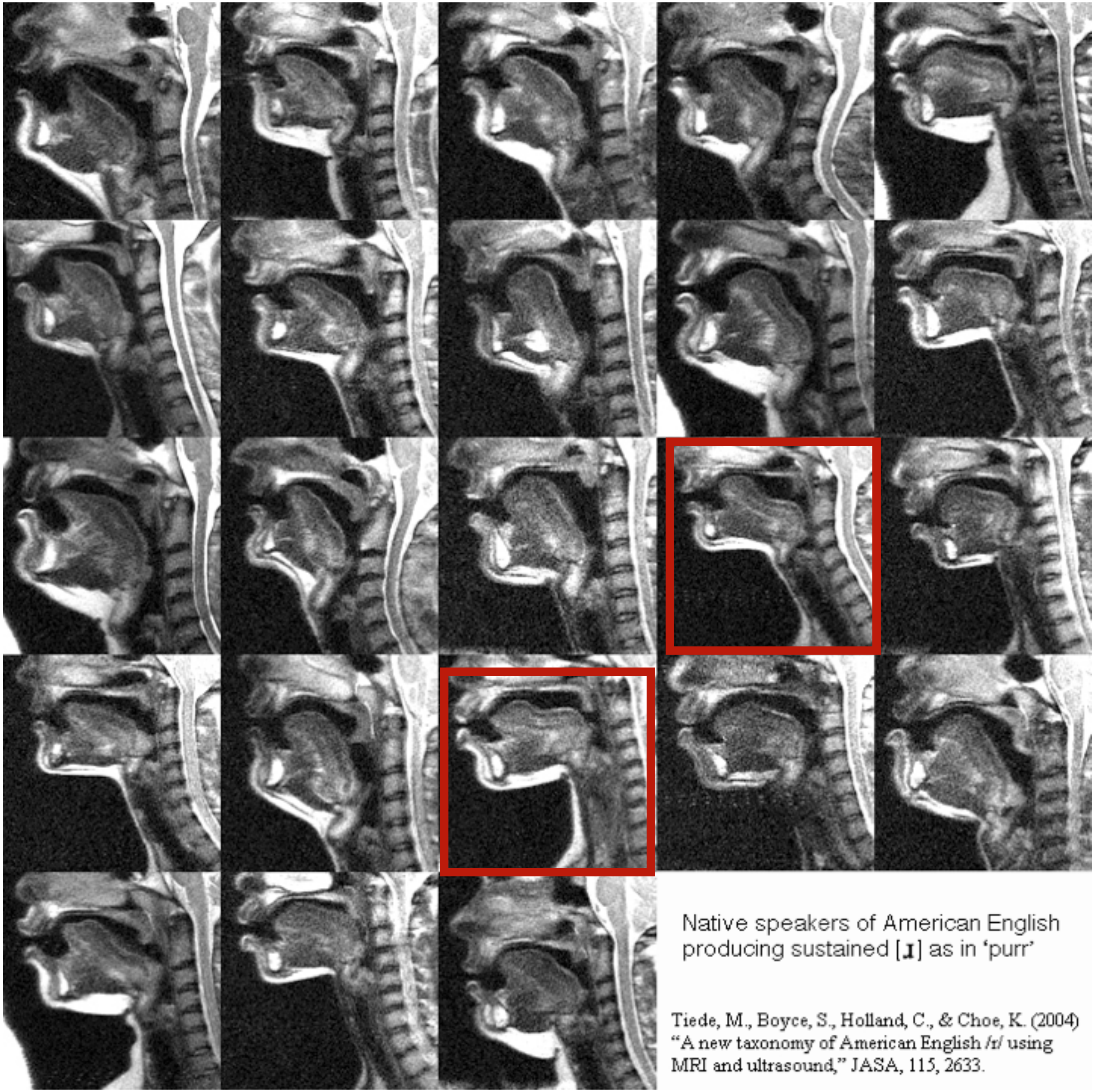
- Tongue Tip
  CD: stop
- Tongue Body
  CD: approximant, CL: uvular



"rye"

- Lips
  CD: approximant
- Tongue Tip/Body
  CD: approximant, CL:palatal
- Tongue Root
  CD: approximant





Native speakers of American English producing sustained [ɹ] as in 'purr'

Tiede, M., Boyce, S., Holland, C., & Choe, K. (2004) "A new taxonomy of American English /r/ using MRI and ultrasound," JASA, 115, 2633.

Traditional (IPA) description of consonants and gestural analysis

(1) Laryngeal gesture results:

- voiced (<laryngeal narrowing)
  voiceless (<laryngeal opening)

(2) Location of oral constriction gesture

- bilabial, labiodental
  dental, alveolar, palato-alveolar
  palatal, velar, uvular, pharyngeal

(3) central or lateral

(4) Velic gesture results:

- nasal (<velic opening)
  oral (<velic closure)

(5) Degree of oral constriction gesture

- stop
  fricative
  approximant

Vowel and consonant gestures

- How do vowel gestures differ from consonant gestures?

(1) consonants are more constricted than vowels

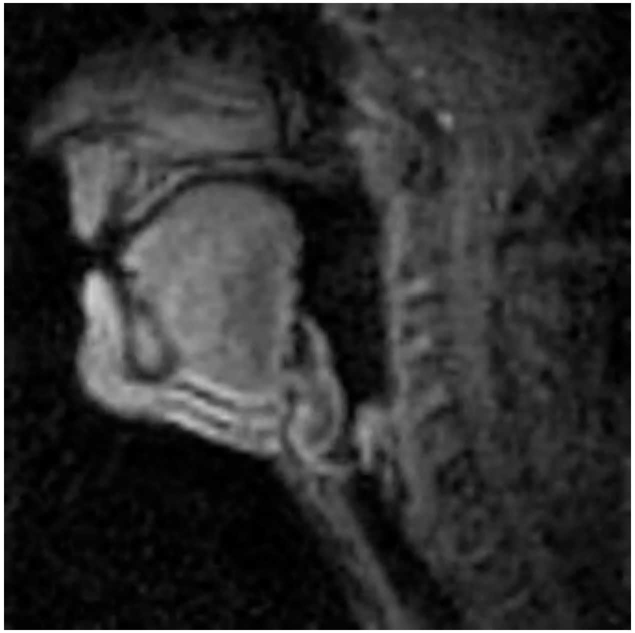
  - exceptions?

(2) vowel gestures are formed more slowly and "last longer" than consonant gestures

Vowel gestures

3 basic vowels that occur in almost all languages

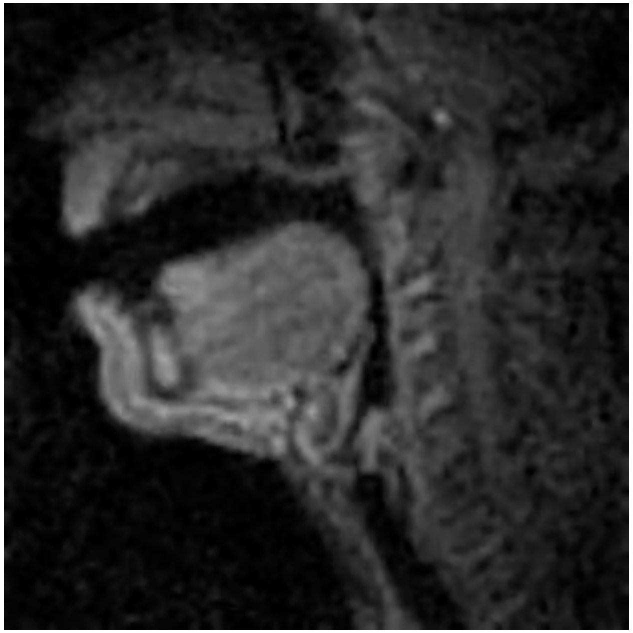
“heed”



TONGUE BODY

TBCL = PALATAL

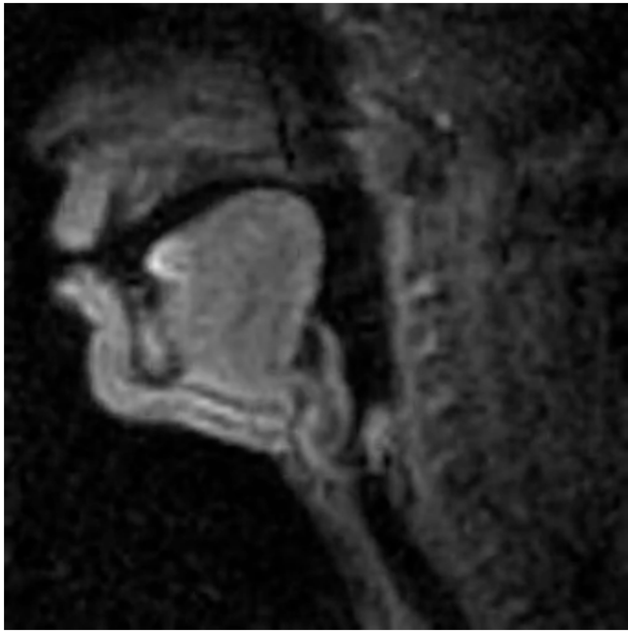
“hod”



TONGUE ROOT

TBCL = PHARYNGEAL

“who’d”



LIPS + TONGUE

TBCL = VELAR

Systems for representing vowels

Location and degree of dorsal constriction

degrees: narrow mid wide (and intermediate)
locations: palatal velar uvular pharyngeal

Tongue Position system

high-low position of tongue body
front-back position of tongue body
round-unrounded

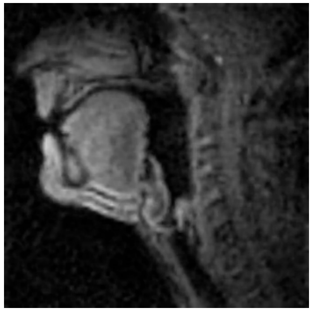
Formant (resonance) system

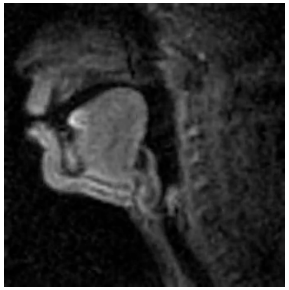
- high-low value of F1 (related to tongue body height)
- high-low value of F2 (related to tongue body front-back)

FRONT BACK

HIGH

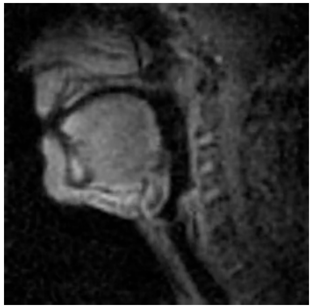
“heed”

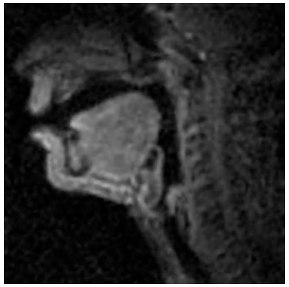




“who’d”

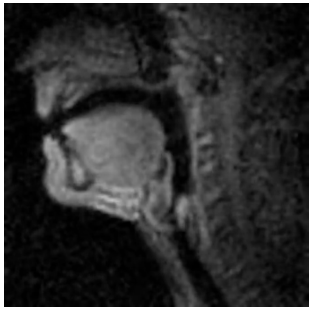
“hid”

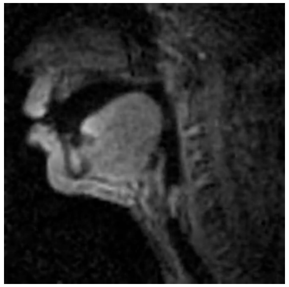




“hood”

“head”

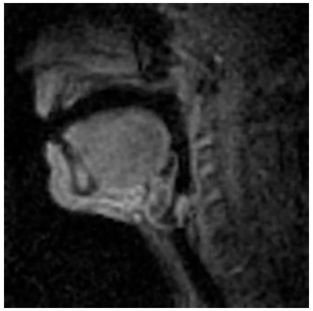


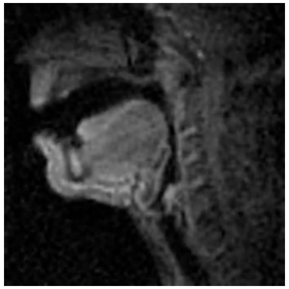


“hoed”

LOW

“had”





“hod”

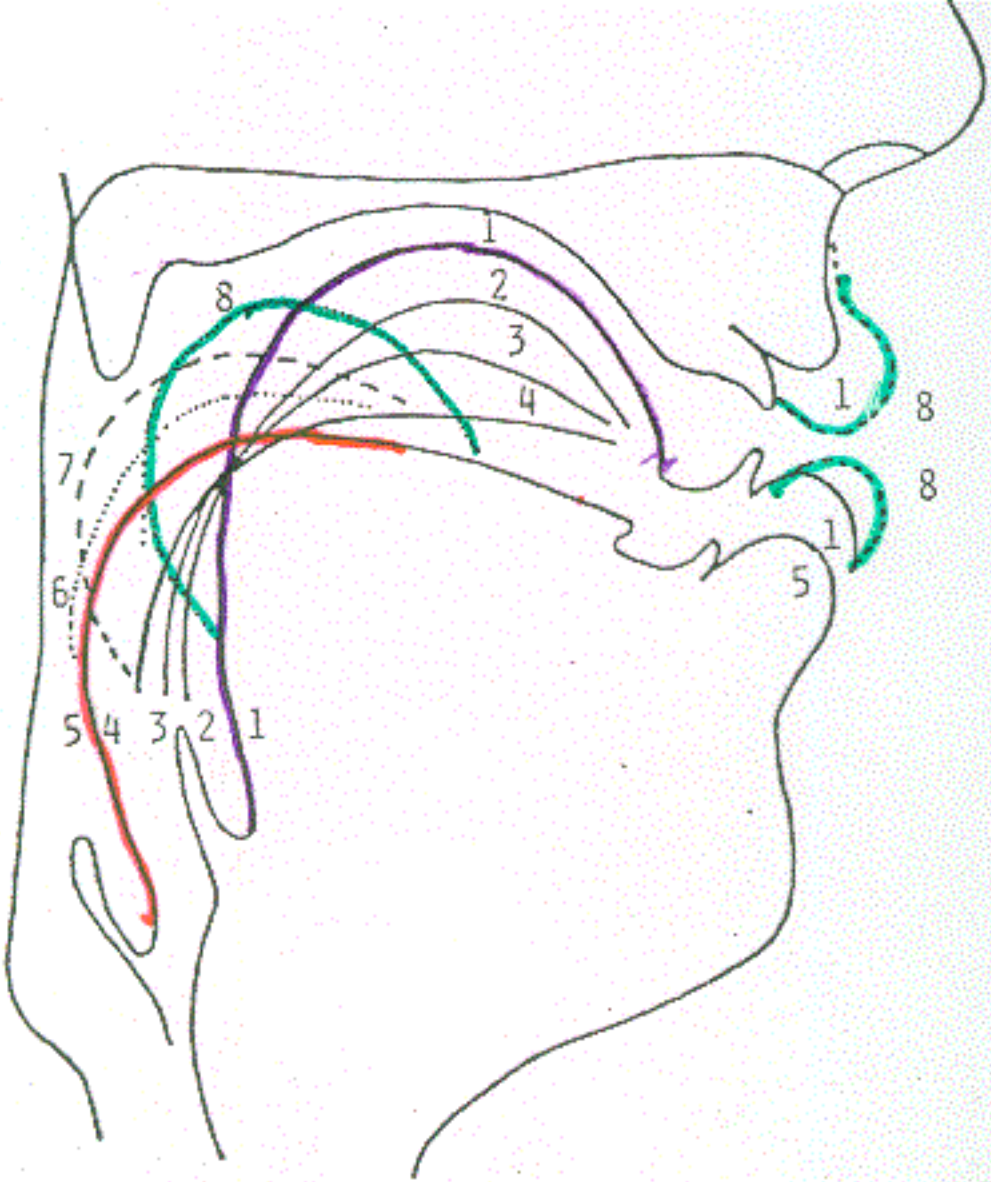


Fig. 1.4 The positions of the vocal tract in the author´s pronunciation of the vowels in the words: (1) "heed", (2) "hid", (3) "head", (4) "had", (5) "hod", (6) "hawed", (7) "hood", and (8) "who´d".