Tones in Cantonese:
Articulatory vs. Acoustic Representation

IACL18 / NACCL22
Harvard University
May 20-22, 2010

Jon Nissenbaum
Dept. of Linguistics
McGill University

Acknowledgements:

My collaborators at the University of Montreal and the Massachusetts Eye and Ear Infirmary:

Gilles Beaudoin, **Guillaume Gilbert,**
Hôpital Notre-Dame/University of Montreal

Jennifer Kan, *MIT*

John E. Kirsch, *Siemens Medical Systems*

James B. Kobler, Hugh D. Curtin, Robert E. Hillman,
Massachusetts Eye and Ear Infirmary/Harvard Medical School

Introduction:

- Study of articulatory gestures has aided in the development of phonetic and phonological theories
- Detailed knowledge exists concerning how gestures of the tongue body and blade, lips and velum interact in speech production
- But the gestures of the larynx and vocal folds that control pitch (f0) dynamically during speech remain poorly understood

Questions and Goals

1. Is phonological tone properly characterized in acoustic or articulatory terms (and does it matter)?

   --> Are phonological ***register*** and ***tone*** features subserved by distinct articulatory mechanisms?

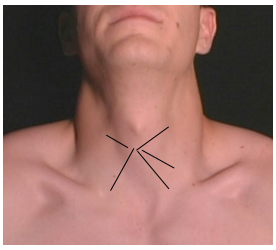
> ***Goal 1:*** ***Report evidence from two studies indicating that register and tone in Cantonese have distinct physiological control systems***

2. How do you make a movie of laryngeal articulations with a frame rate of 100 fps...

   ...when the camera's "shutter speed" is just ***1*** fps?

> ***Goal 2:*** ***Generating image sequences with a fast enough frame rate to image the larynx & vocal tract dynamically***

- Our ignorance about laryngeal gestures and their relation to phonological categories is due in part to the inaccessibility of the larynx
- Laryngeal movements can be (and have been) studied by measuring electrical potentials generated by muscles during speech production (EMG)
- But EMG requires insertion of electrodes into the neck
  … and there are half a dozen pairs of intrinsic laryngeal muscles that play critical roles in speech



What about non-invasive imaging techniques like MRI?

The articulations underlying speech sounds are too fast for MRI

- Speech gestures:
  - transition from one target position to another is typically 80–300 ms for a single articulator
  - some consonant articulations can span 20-40 ms [Stevens 1998:38–48]
  - the problem is compounded by coordination of temporally overlapping gestures
  - ***Temporal resolution of 50–100 images/sec (10–20 ms per image) is needed***
- MRI:
  - typically requires at least ***1 sec*** for each image
  - ***Too slow by a factor of 50-100***

A technique for 'freezing' motion using MRI

- *Allows imaging at 144 frames per second*
- *subjects repeat a movement 256 times, every 2 seconds resulting in an "averaged" movie of a single repetition*

Hai^{5} fu^{1}

"It's skin"

High level/falling tone on /fu^{1}/ produced by male Cantonese speaker

The characterization of tone contrasts

- Cantonese provides a useful testing ground for theories of tone representation because of its rich tonal inventory
  - Six-way phonological contrast

The six tones divide into two registers:

Upper register	***high level / falling*** ① *fù* *"skin"* ***mid-high level*** ③ *fu* *"wealthy"*	***high rising*** ② *fú* *"tiger"*
Lower register	***mid-low level*** ⑥ *fuh* *"father"* ***low level / falling*** ④ *fùh* *"to hold on"*	***low rising*** ⑤ *fúh* *"woman"*

Study 1

Evidence from pitch tracks of tones read in a carrier sentence

Experimental Procedure:

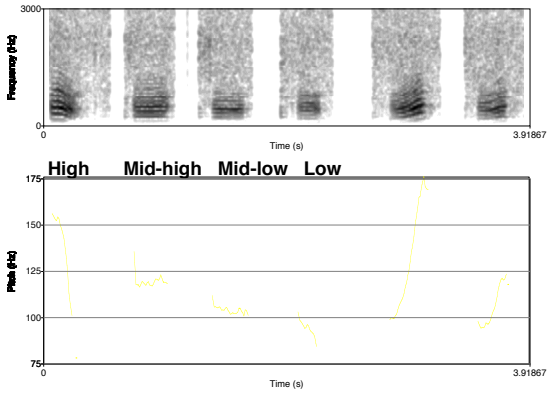
A list of syllables with exponents for all six tones was selected with the aid of a Cantonese dictionary, and in consultation with native speakers.

- *For each set of six items, subjects first read the list of words in their isolation forms.*
- *Each target word was then embedded in a short carrier sentence (a question-answer dialogue), which subjects read aloud.*
- *f0 tracks were extracted from the isolation forms (using Praat) and compared with f0 tracks of the same words in carrier sentences.*

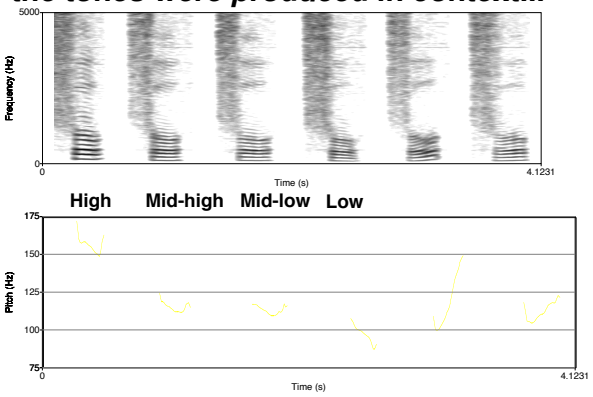
Subjects:

Eight subjects participated (5 male, 3 female). Subjects were all native speakers of Cantonese between the ages of 28 and 48.

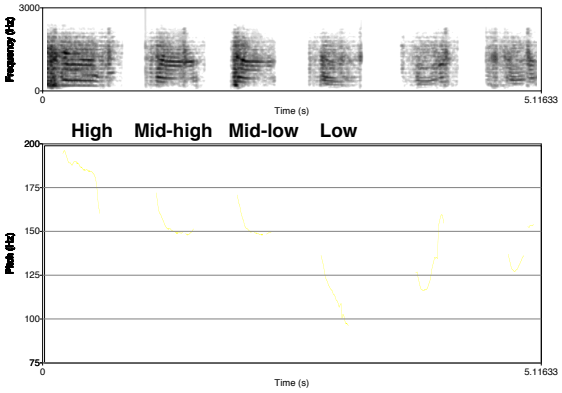
In citation form, the four basic tones divide acoustic space of F0 into four roughly equal regions



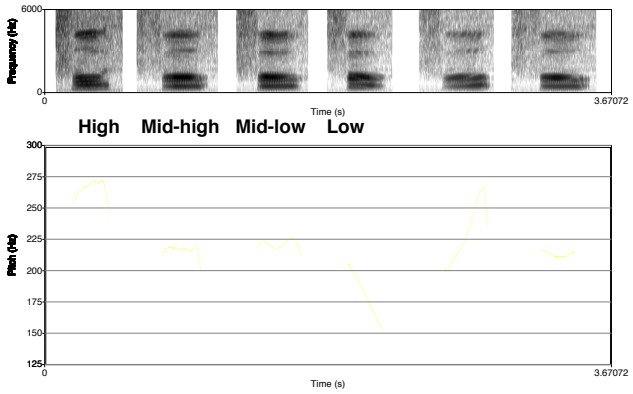
A striking difference emerged when the tones were produced in context...



The acoustic space becomes divided unevenly…
With the two mid-tones at essentially the same F0



This pattern was observed with remarkable consistency across speakers (5 male, 6 female)



Summary table:

f0 measured at 30 ms after onset of target vowel

Tone	*Male speakers*					*Female speakers*		
1	161	154	167	157	169	190	264	225
3	114	103	121	128	129	149	218	201
6	110	106	120	123	125	146	222	196
4	98	92	109	108	101	123	197	182

- *The level tones divide f0 space into **three** regions, with tones 3 and 6 forming a single mid-frequency*

Why is F0 space divided this way in Cantonese speakers' production of tones?

=> Could it be that there is really just one mid-tone?

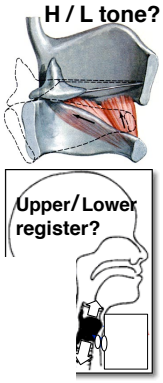
No:

- *A number of lexical generalizations are defined over register, with the "low" mid-tone patterning together with low fall and low rise.*
- *Speakers intuit a difference, which emerges in the citation forms*

An alternative explanation

The four basic tones of Cantonese are represented in terms of two independent articulatory parameters

Two mechanisms for adjusting vocal fold stiffness



- 1.) Stretching / shortening
  - stretching increases the mechanical stiffness, yielding **higher rate of vibration (f0)**
  - shortening yields **lower f0**
- 2.) Raising / lowering
  - lowering slackens the tissue, yielding **lower f0**
  - raising yields **higher f0**

Predictions:

- 1.) Stretching / shortening
  - **Should be employed to raise / lower the pitch *independent of pitch register***
- 2.) Raising / lowering
  - **Should be employed to change register *independent of tone level within that register***

Study 2

Evidence from magnetic resonance images (MRI) of the vocal folds during production of tones

Description of methods

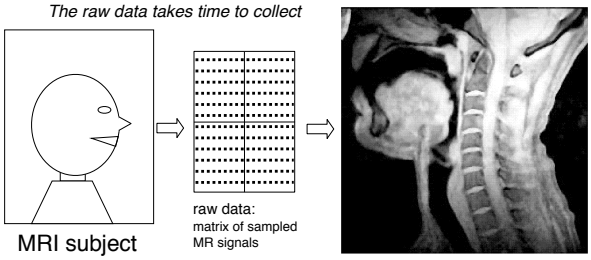
Imaging articulatory gestures of the larynx

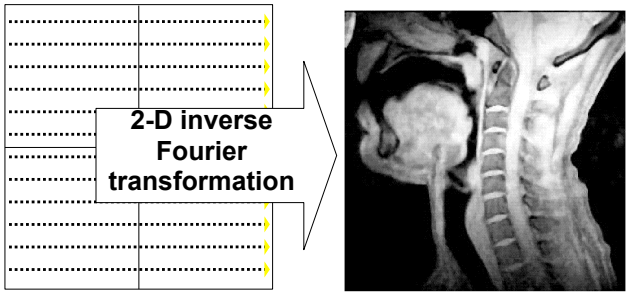
- *How is it possible to create a high-speed movie with a camera whose shutter speed is so slow?*
- In order to understand how to get around the "slow shutter speed" of MRI, it is important to see exactly what happens while the shutter is open

Why is the time resolution of MRI so poor?

- The MR image is not generated by exposing film
- The relation between the subject and the image is mediated by a matrix of sampled magnetic resonance signals, whose frequencies and phases encode the spatial location of resonating protons inside the subject

The raw data takes time to collect





- ***Image is then constructed automatically from the raw data***

✦ *The matrix is filled incrementally, row by row, in 256 discrete steps*

At 7 msec increments, it takes roughly 1.8 seconds to sample 256 rows of raw data

How do we "beat the clock" to create a fast image sequence?

- Answer:

  Take advantage of the fact that during the 1.8 seconds that the MRI "shutter" is open, there are really *256 discrete events ==> each lasting just 7 milliseconds.*

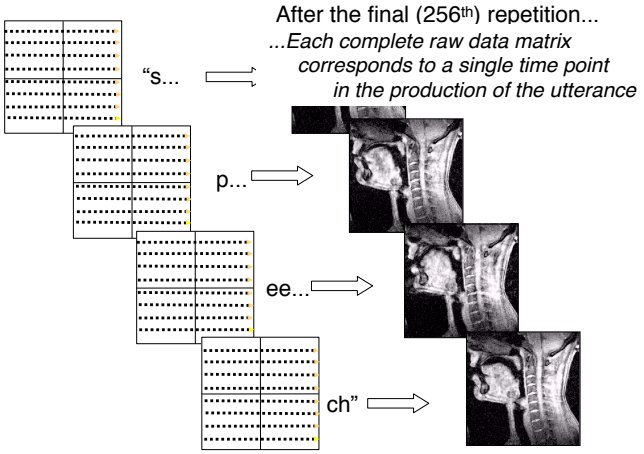
- To create a high-speed movie, the raw data is collected in an unconventional order.
- *Each row of raw data is sampled multiple times while the subject repeats a short utterance over and over.*

  (Masaki et al. 1997, Mathiak et al. 2000, Mohammad et al. 1997)

- In summary:
- The phases of a movement can be, in effect, "frozen in time" long enough to be scanned -- by repeating it and taking a *very partial* scan each time.

In the end there are *multiple* instances of each line of raw data

- *That means:*
  - There is a complete raw data set corresponding to *each point in time* relative to the speech onset
  - Each raw data set yields an image
  - The result is a set of images in sequence (each composed of data from multiple repetitions), incremented at 7 ms intervals
    - There are 144 such intervals per second



Experimental Procedure:
Preliminaries

A list of syllables with exponents for all six tones was selected with the aid of a Cantonese dictionary, and in consultation with native speakers.

A single item from the syllabary, fu*, was chosen for the MRI experiment, since the MRI protocol required 256 repetitions at 2 second intervals (= 8.5 minutes) for each of the six items.*

Fu *has widely recognizable meanings for all six tones; it is an open syllable with a simple (monophthongal) vowel, thus minimizing interfering articulatory activity.*

In order to ensure that the target word remained meaningful for the subject throughout the 256 repetitions, a short question-answer dialogue was presented through MR-compatible headphones.

Prior to each scanning session, subjects read the the question-answer pairs.

Format for the question-answer pairs:

Q: "_____ or _____ ?"
(contrasting word) (target word, fu*)*

A: "It's _____ ."
(target word, fu*)*

Subject hears question-answer pairs through headphones and repeats the answer part along with the tape during the MRI scan. Repetition interval is 2 seconds.

Experimental Materials

	level / falling	**mid-level**	**rising**
Upper register	*Gwàt waahk jé fù?* Bones or skin? *Háih fù.* It's skin.	*Kùhng waahk jé fu?* Poor or wealthy? *Háih fu.* It's wealthy.	*Lùhng waahk jé fú?* Dragon or tiger? *Háih fú.* It's a tiger.
Lower register	*Tuhy waahk jé fùh?* To push or to hold on? *Háih fùh.* It's "to hold on."	*Moúh waahk jé fuh?* Mother or father? *Háih fuh.* It's the father.	*Naàhm waahk jé fúh?* Man or woman? *Háih fúh.* It's a woman.

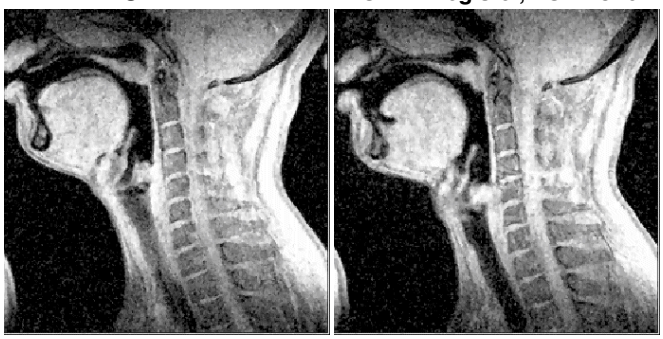
Subjects:

Nine subjects participated (6 male, 3 female). Subjects were all native speakers of Cantonese between the ages of 19 and 54.

Preliminary results from the MRI study

Male speaker age 20 **Upper and Lower extreme tones**

UPPER register, HIGH tone **LOWER register, LOW tone**



Mid tones

UPPER register, L-tone ***("high mid" tone 3)*** **LOWER register H-tone** ***("low mid" tone 6)***

Mid tones

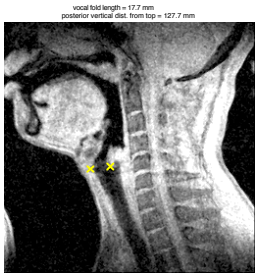
Frame comparison, onset of /u^{3}/ UPPER mid-tone

Mid tones

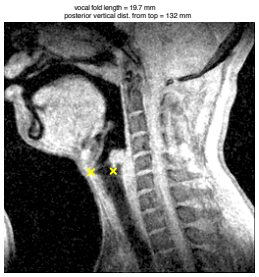
Frame comparison, onset of /u^{6}/ LOWER mid-tone

Mid tones

/u^{3}/ UPPER mid-tone



/u^{6}/ LOWER mid-tone



- Difference in vocal fold length between tones 3 and 6 for this subject is 2 mm (i.e. vocal folds are 11% longer at onset of tone 6 than at onset of tone 3)
- Difference in vertical position is 4.3 mm (i.e. larynx lowers by nearly 1/2 cm for tone 6)

Summary of preliminary findings:

- Tone extremes (High / Upper and Low / Lower):
  - For all subjects, lengthening and raising are observed for the highest tones, lowering and shortening for the lowest tones
- Mid-tones:
  - Comparative data available from fewer subjects
  - For the two subjects with usable images for both mid-tones:
    - the vocal folds undergo lengthening and the larynx is relatively lower for the lower-register mid-tone (tone 6);
    - the vocal folds shorten and raise for the upper-register mid-tone (tone 3)

Conclusions

- Two production mechanisms are known to be available in principle for adjusting pitch independently
- Characterizing tone and register features in terms of these mechanisms does better justice to the facts of Cantonese speech
- The articulatory model is supported by high-speed MRI of Cantonese speakers' production of tones.

Conclusions

- It has proven fruitful in many cases to characterize aspects of phonological knowledge in terms of the articulatory dimension of speech.
- When it comes to intonation (and other aspects of sound structure under the control of the larynx), linguists have tended overwhelmingly to characterize theoretical primitives in terms of acoustic dimensions (f0, VOT, etc.)
- These results suggest that there is something to be gained by taking the articulatory dimension of tone to be theoretically primary
- A non-invasive technique for investigating aspects of speech production that have resisted empirical study, potentially making a broad range of questions available for novel methods of inquiry.