

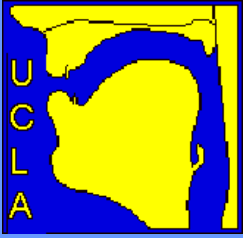
# Linguistic Voice Quality

Patricia Keating

University of California, Los Angeles

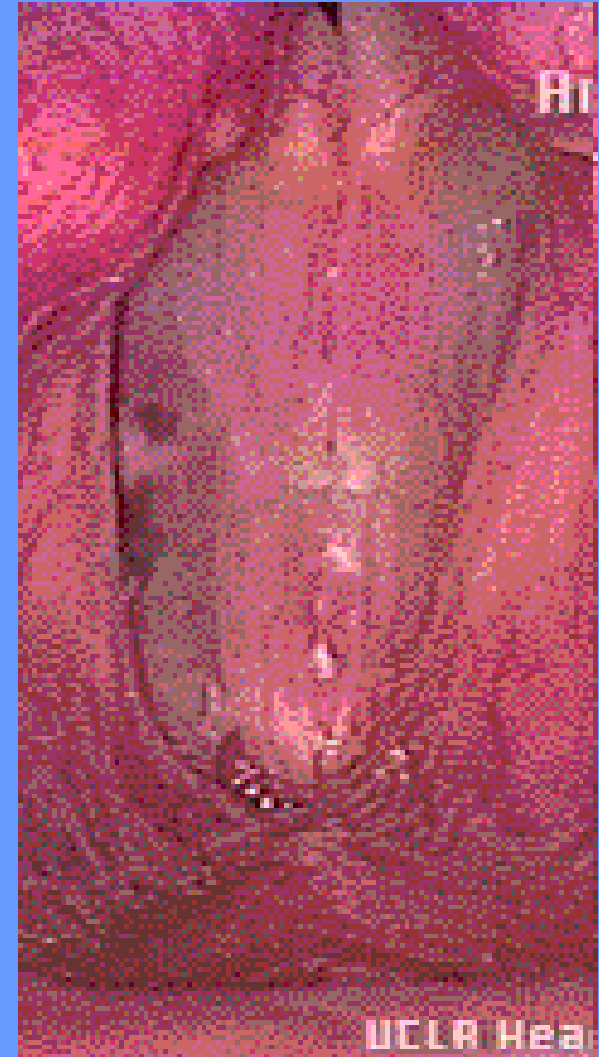
Christina Esposito

Macalester College, St. Paul



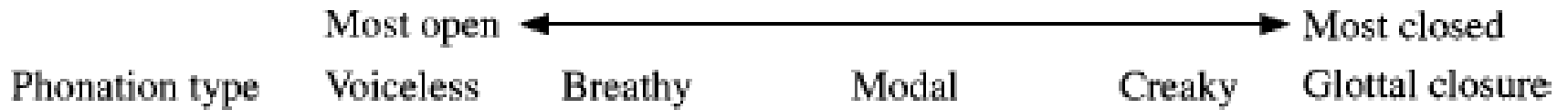
# Phonation

- Production of sound by vibration of the vocal folds
- **Phonation type contrasts** on vowels and/or consonants

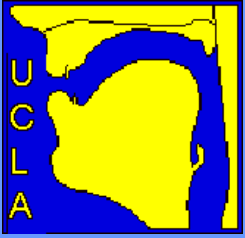




# Ladefoged's (simplified) glottal constriction model

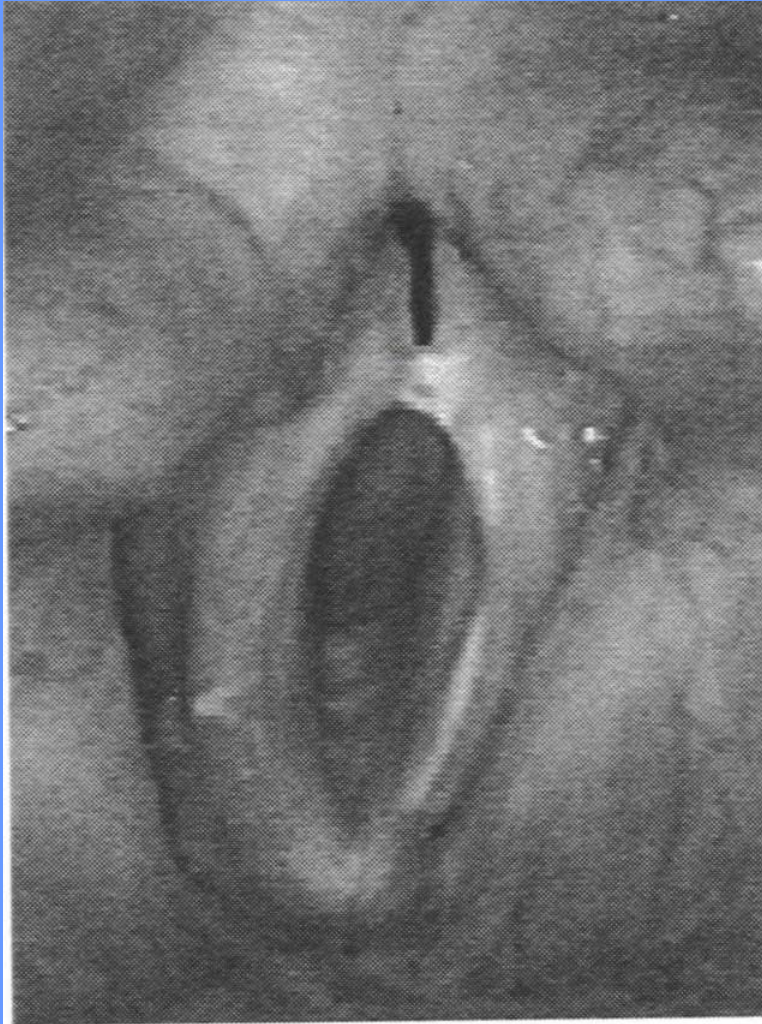


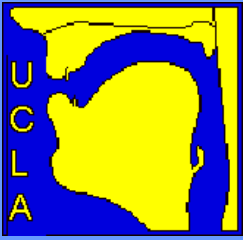
- Size of glottal opening varies from that for voiceless sounds (no phonation) to zero (glottal closure)
- Phonation is possible at a variety of constrictions, but with voice quality differences
- These are the most common contrasts



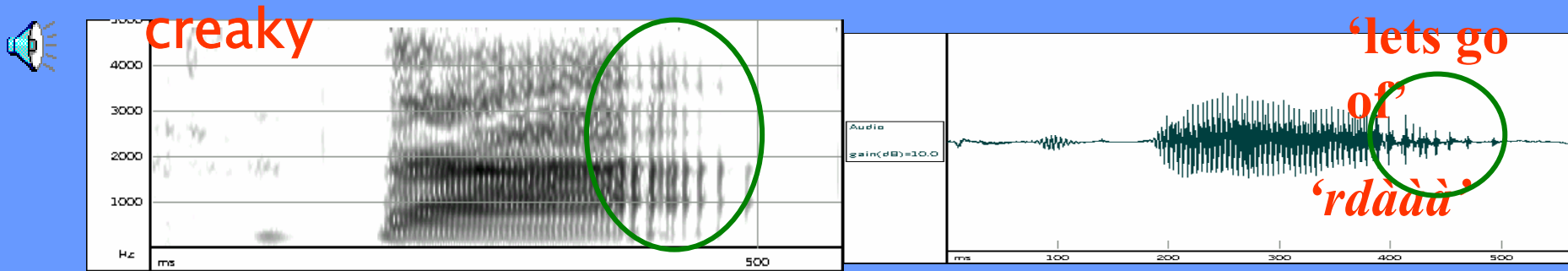
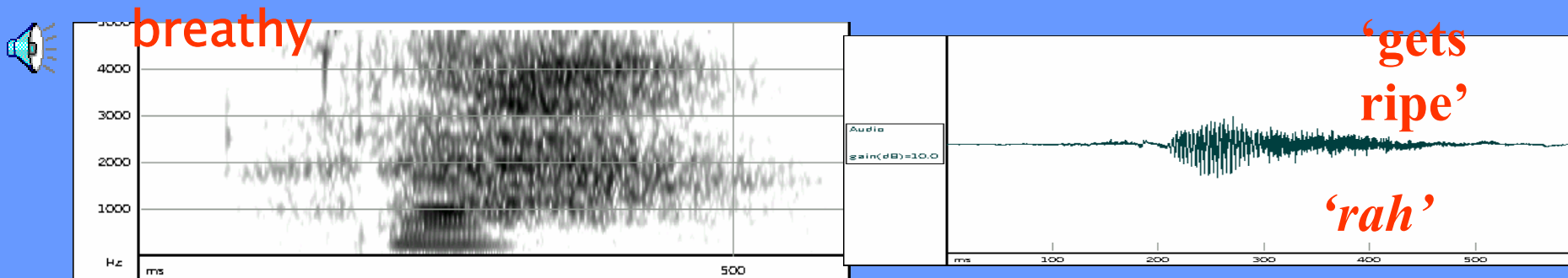
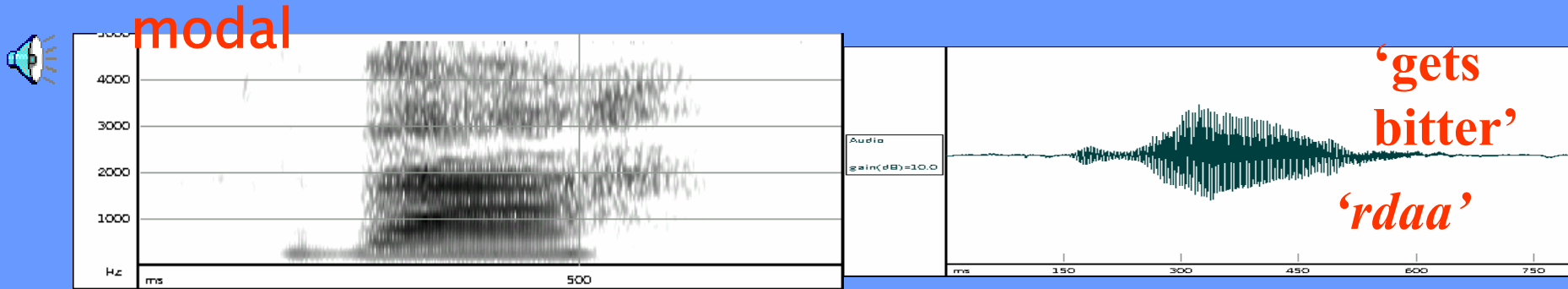
# Breathy vs. creaky glottis

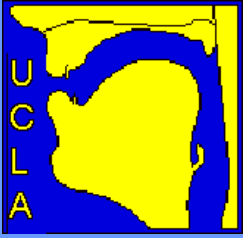
(from Ladefoged)



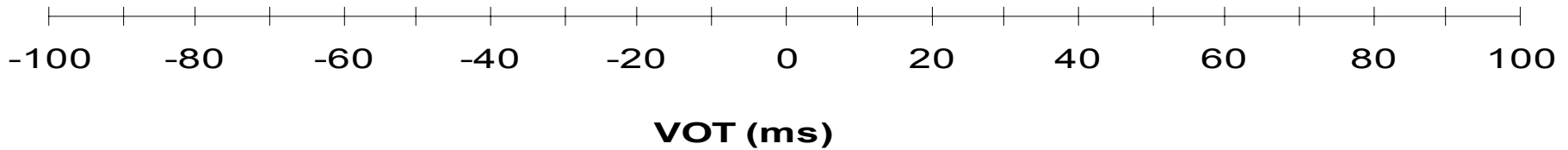
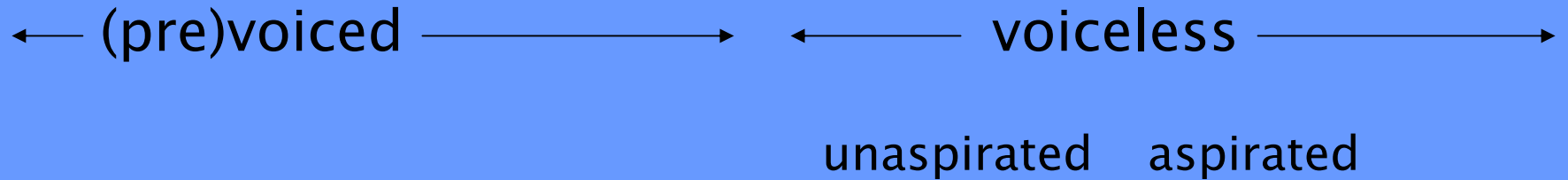


# 3 phonations of San Lucas Quiavini Zapotec





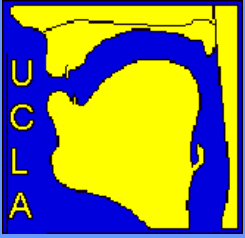
# A continuum like VOT



lead  
VOT

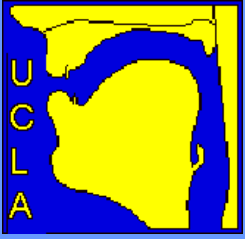
short  
lag  
VOT

long  
lag  
VOT



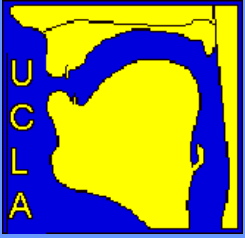
# This talk

- Language differences in acoustic dimensions of phonation contrasts
- Perception of phonation contrasts
- A bit on phonation and tones



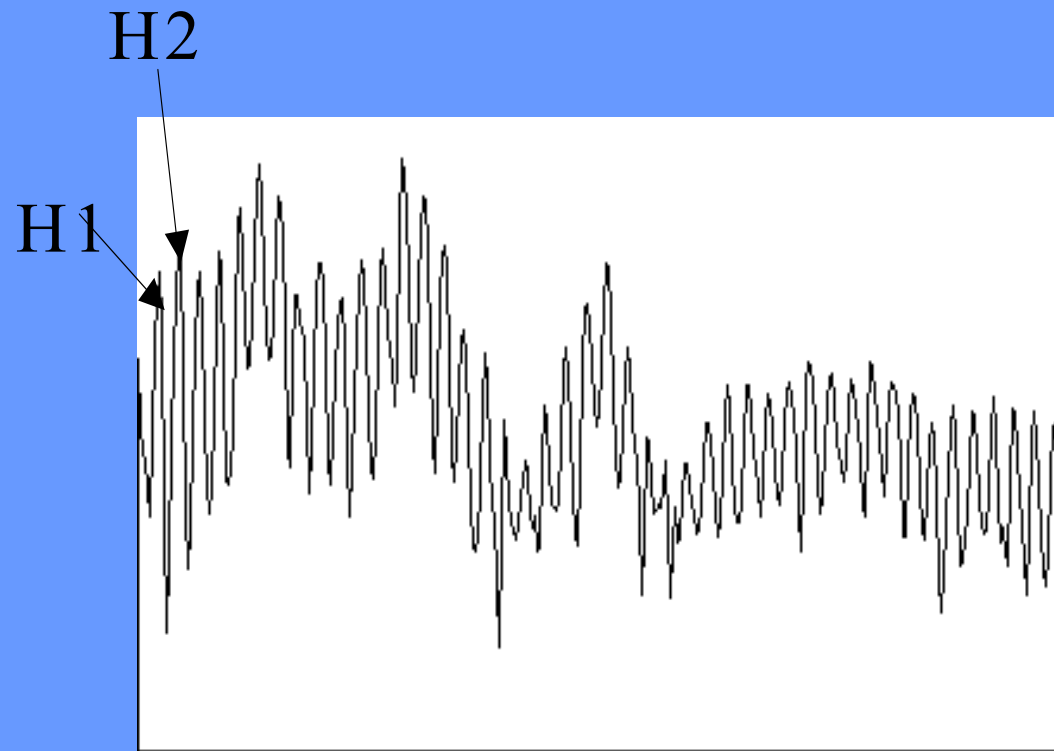
# Spectral measures of voice quality

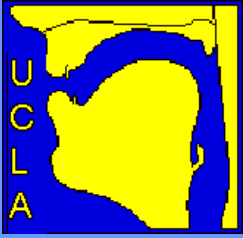
- Given the  $F_0$ , the **frequencies** of all the harmonics are determined and cannot vary
- But the **amplitudes** of the harmonics do not depend on the  $F_0$  and can vary
- **Relative amplitudes** of harmonics can be readily seen in a spectrum



# Most popular measure: H1-H2

- Relative amplitude of first two harmonics H1 and H2
- Breathy voice: strong H1
- Creaky voice: H1 weaker than H2





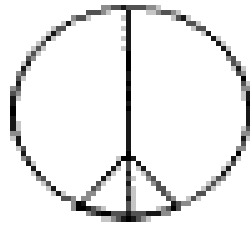
# Related to Open Quotient

- Vocal fold vibration cycle divides into open vs. closed portions
- Open portion of cycle as proportion of total cycle: **Open Quotient (OQ)**
- The more time the vocal folds are open, the more air gets through, so the breathier the voice
- Most extreme OQ would be 1.00: folds don't close completely and are always letting some air through



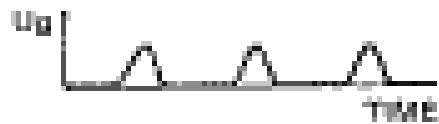
# Glottal constriction and OQ (from Klatt & Klatt 1990)

LARYNGEALIZED



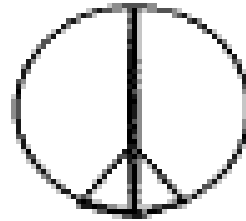
1A

Ug

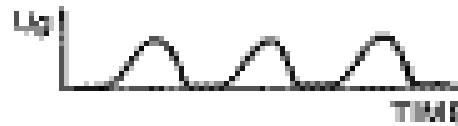


1B

MODAL

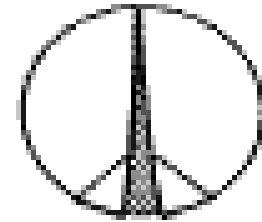


2A

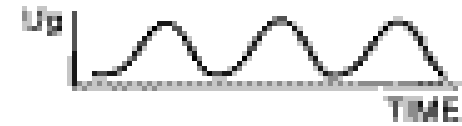


2B

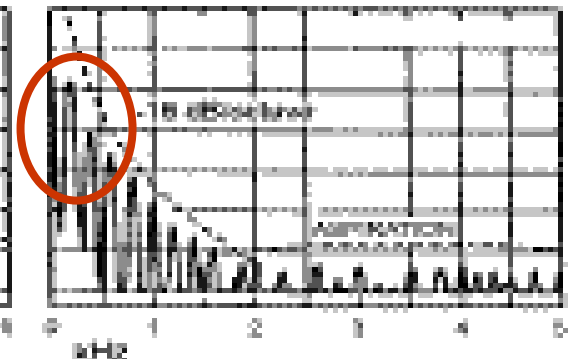
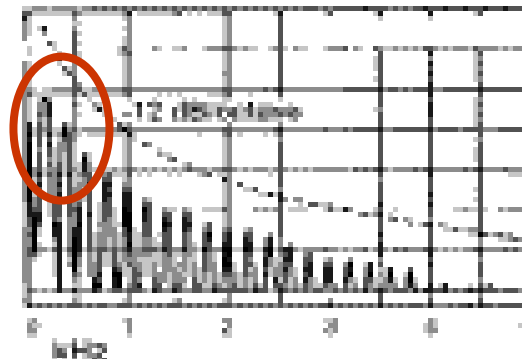
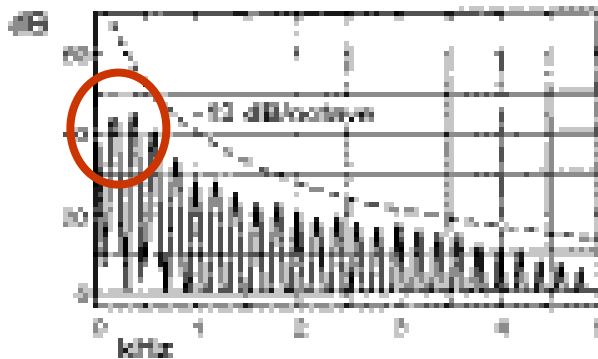
BREATHY



3A

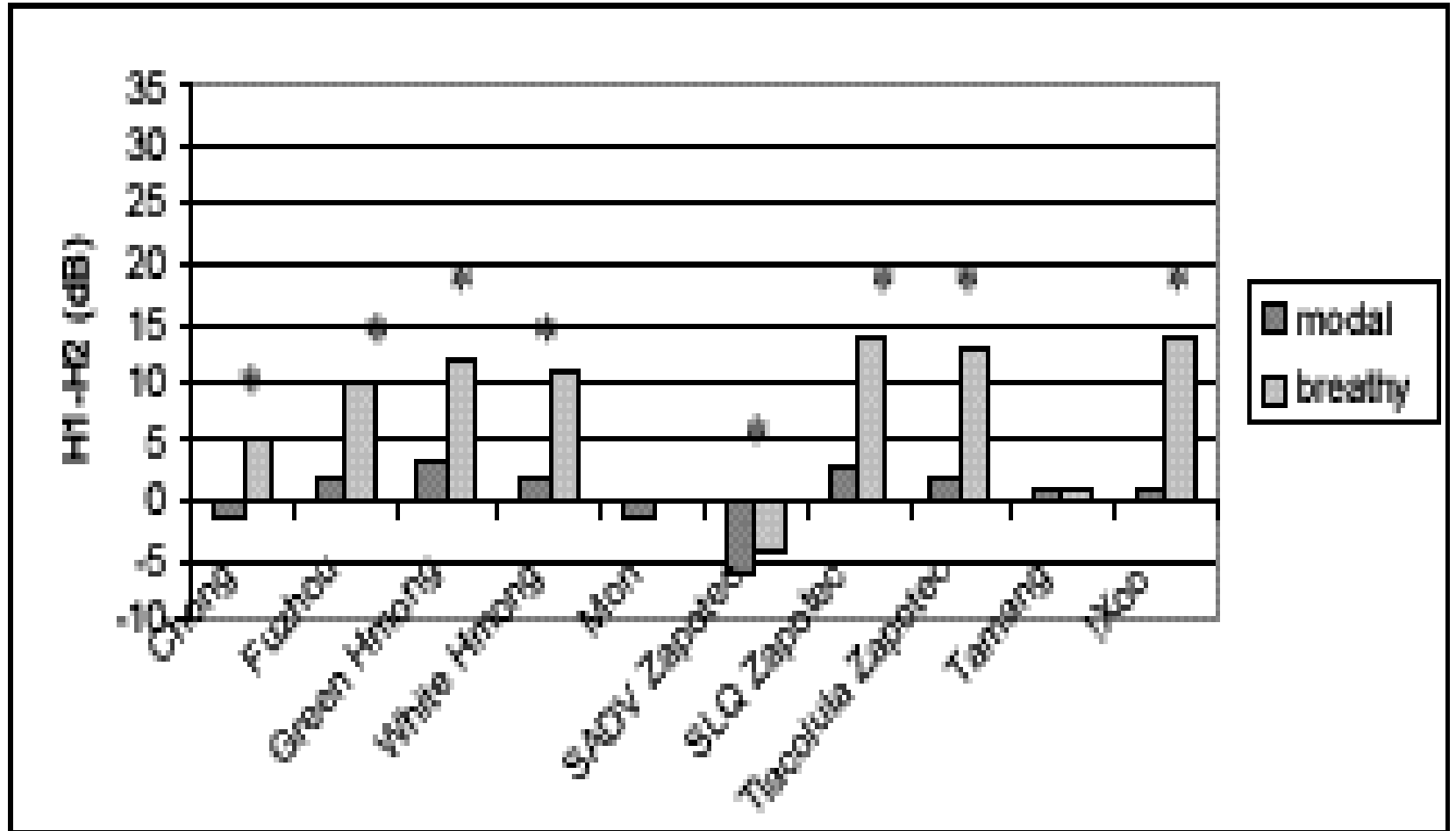


3B

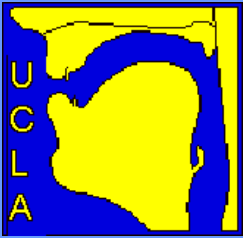




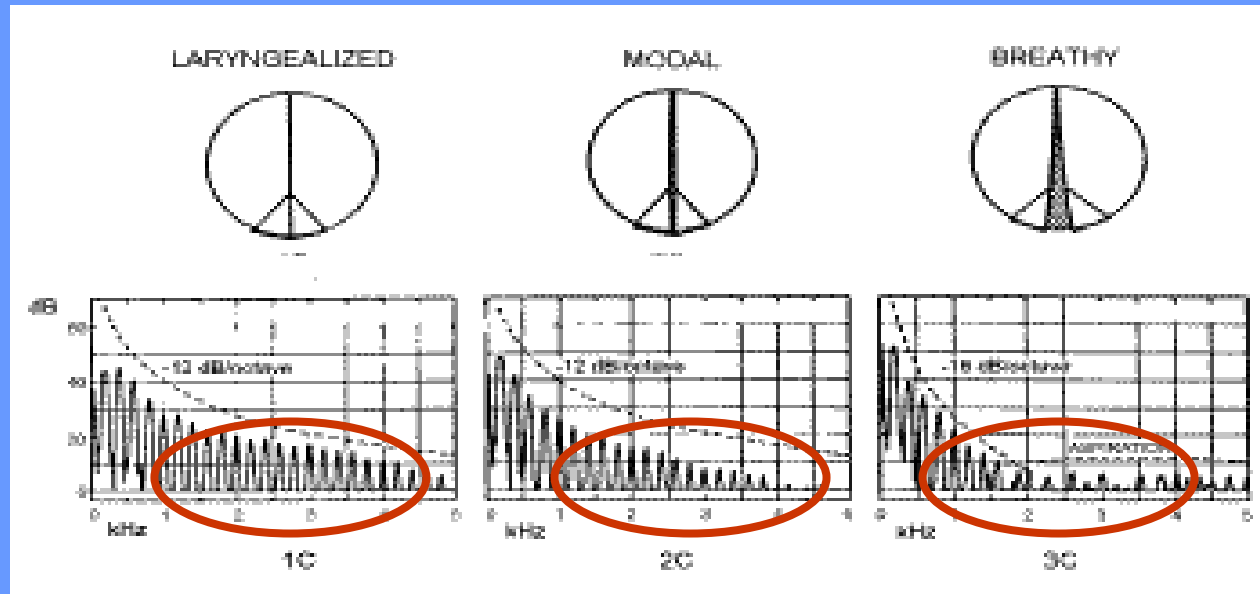
# H1-H2 and breathy voice in several languages



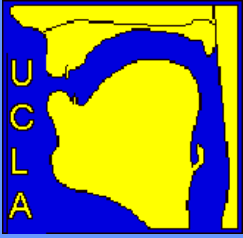
from Esposito 2006



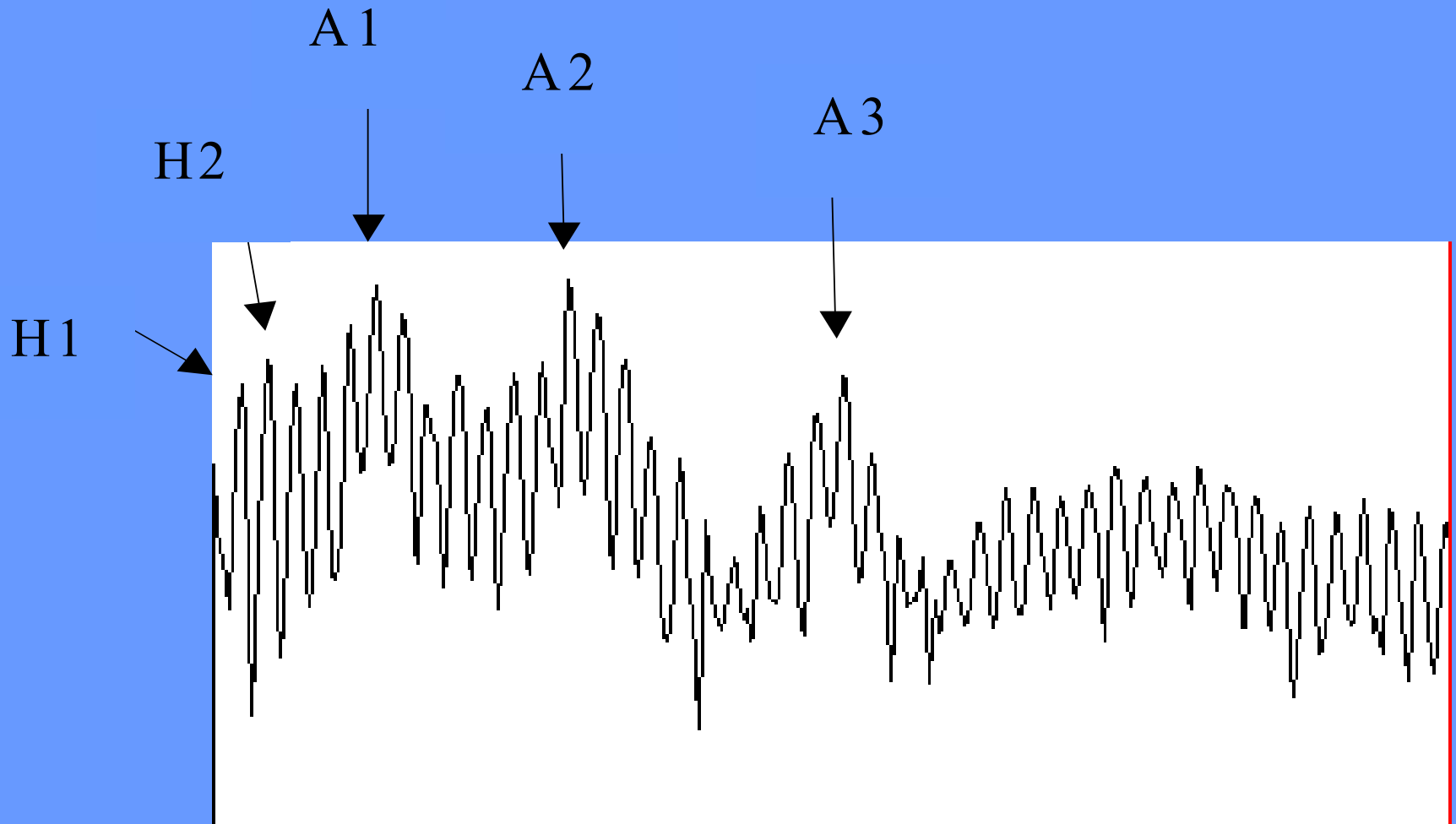
# Spectral tilt differences



- Stronger high frequency components with more abrupt closing of the folds is typical with greater glottal constriction
- Several ways to quantify overall tilt



# H1 and A1, A2, A3



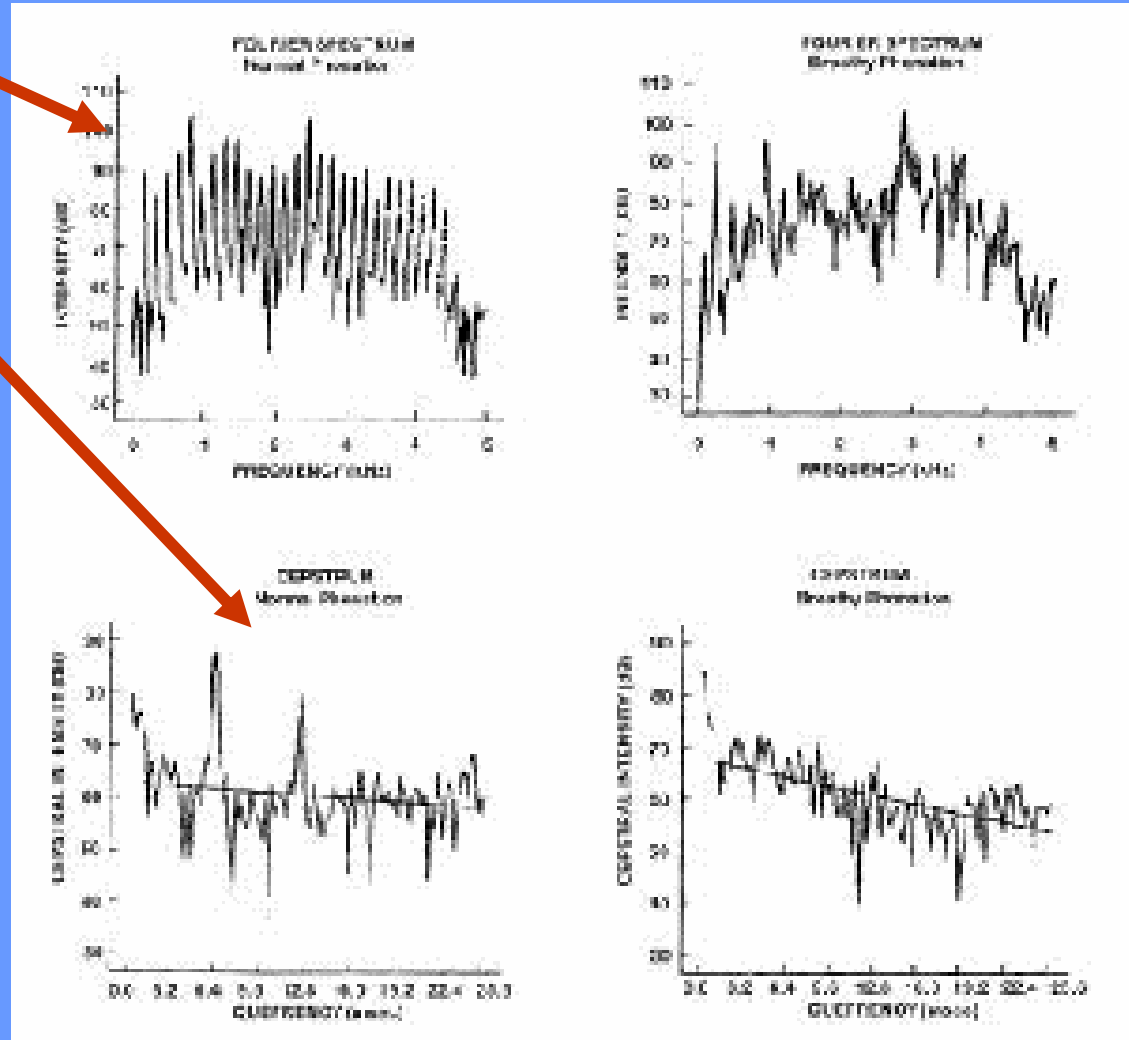


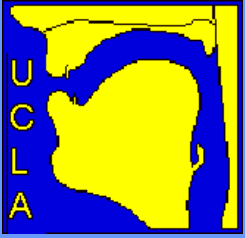


# Cepstral Peak Prominence

(from Hillenbrand et al. 1994)

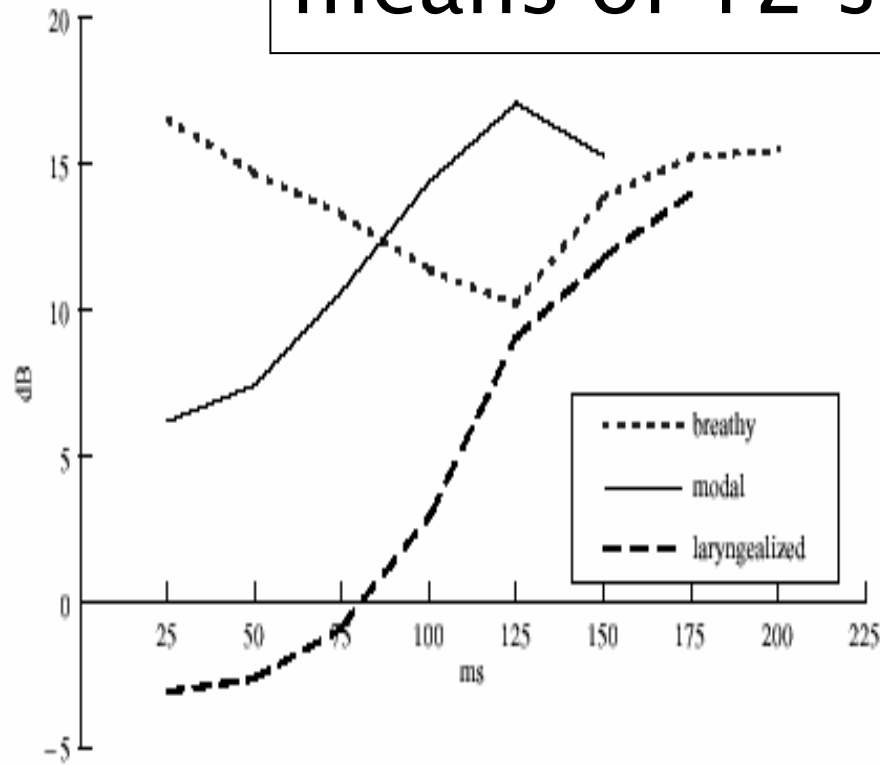
- Well-defined harmonics give strong peak in cepstrum
- Harmonics and cepstral peak less defined in breathy noise (on the right)



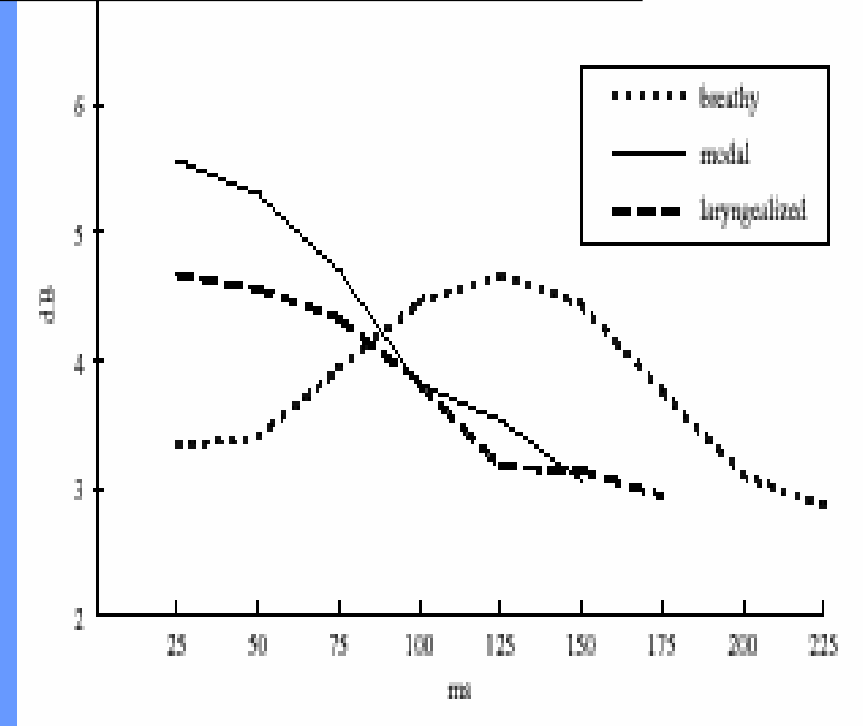


# H1-H2 and CPP in Mazatec (from Blankenship 1997)

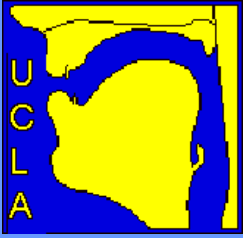
means of 12 speakers x 3 reps



H1-H2



CPP



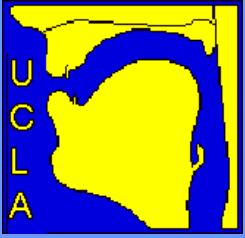
# Summary

## CREAKY

- H1-H2 is low
- Higher frequencies are strong
- Cepstral Peak Prominence can be low due to irregular vibration

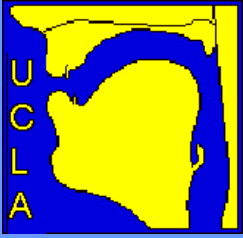
## BREATHY

- H1-H2 is high
- Higher frequencies are weak
- Cepstral Peak Prominence is low due to noise



# Within-language difference in phonations

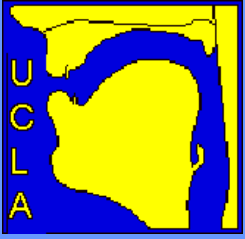
- OQ and tilt measures are generally correlated, but not always
- Contrasts that are not distinguished by H1-H2 are necessarily distinguished by one or more other measures (e.g. H1-A3, H1-A2, CPP in Esposito 2006)
- But even within a language, speakers can differ: Esposito (2003, 2005) on Santa Ana del Valle Zapotec






# Santa Ana del Valle Zapotec

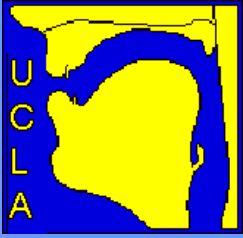
- Spoken in Santa Ana del Valle, Oaxaca, Mexico
- Related to: San Lucas Quiaviní Zapotec, San Juan Guelavía Zapotec, Tlacolula Zapotec



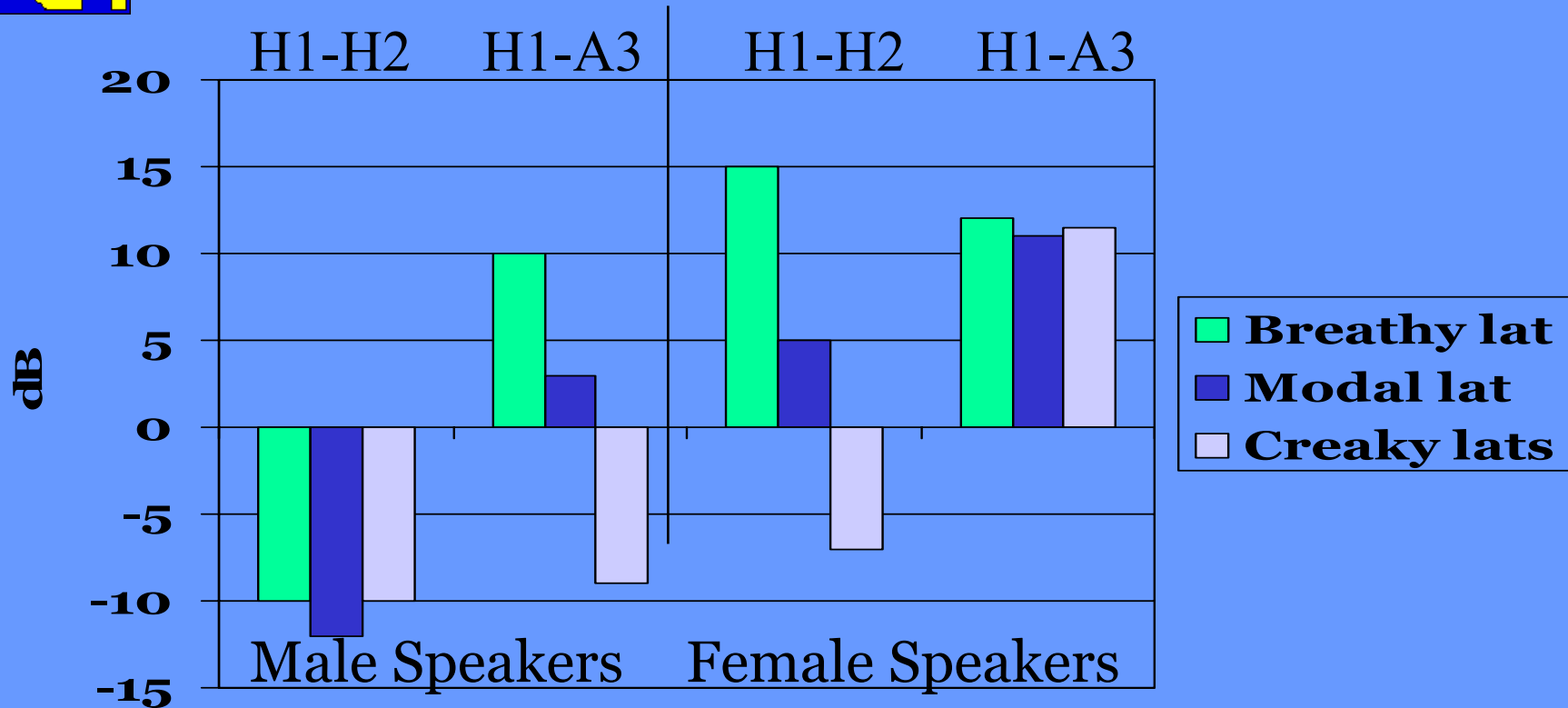


# Santa Ana del Valle Zapotec minimal triple

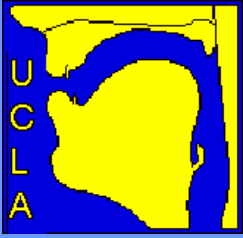
- Modal: ‘can’ lat 
- Breathy: ‘place’ laṭ 
- Creaky: ‘field’ laṭs 



# Spectral measures

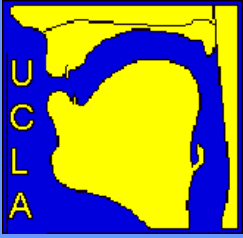


- The three phonations are distinguished by:
  - H1-A3 for the male speakers
  - H1-H2 for the female speakers



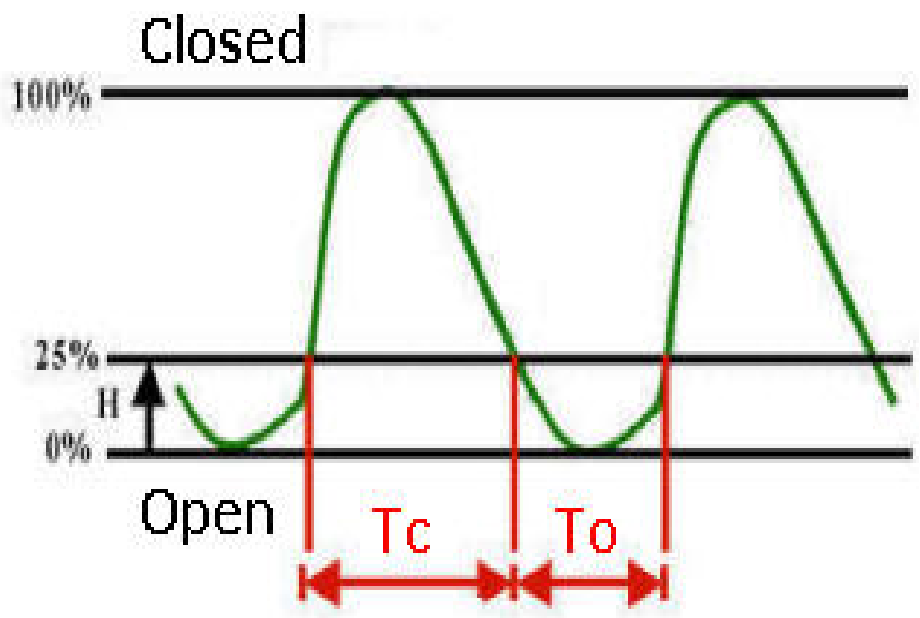
# Compare with EGG

- Are the speakers really producing the contrasts differently as suggested by the acoustic measures?
- Electroglossograph recordings using Glottal Enterprises EGG



# EGG Closing Quotient CQ

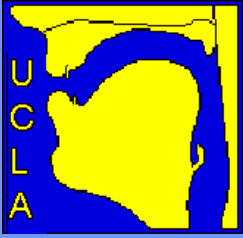
- Reflects the portion of time the vocal folds are closing during each glottal cycle
- Measured automatically



H = 25 %  
threshold

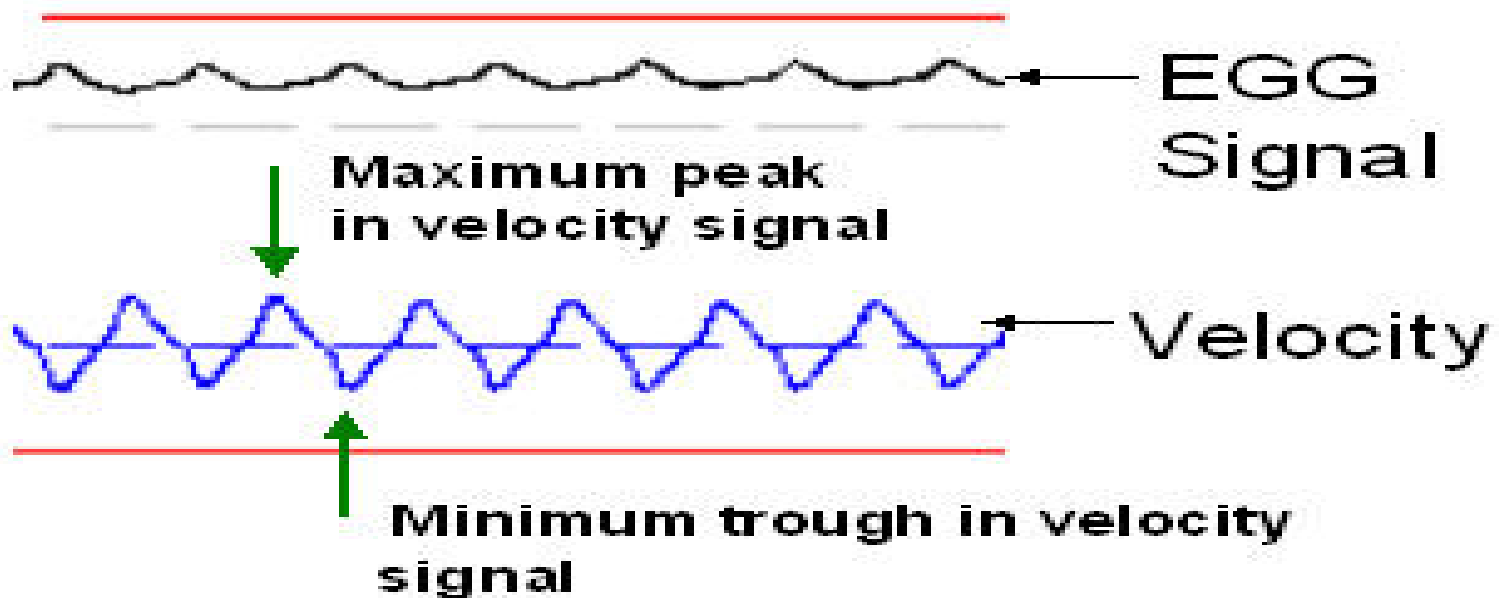
$$CQ = \frac{T_c}{(T_c + T_o)}$$

From <http://www.lpl.univ-lille.fr/~ghio/pedago-EggUK.htm>



# EGG Max-Min Velocity

- A measure of pulse symmetry
- Measured manually from the derivative of the EGG signal

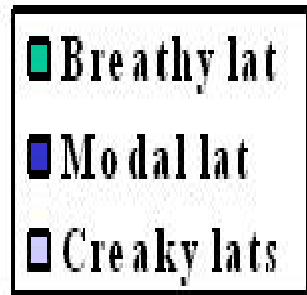
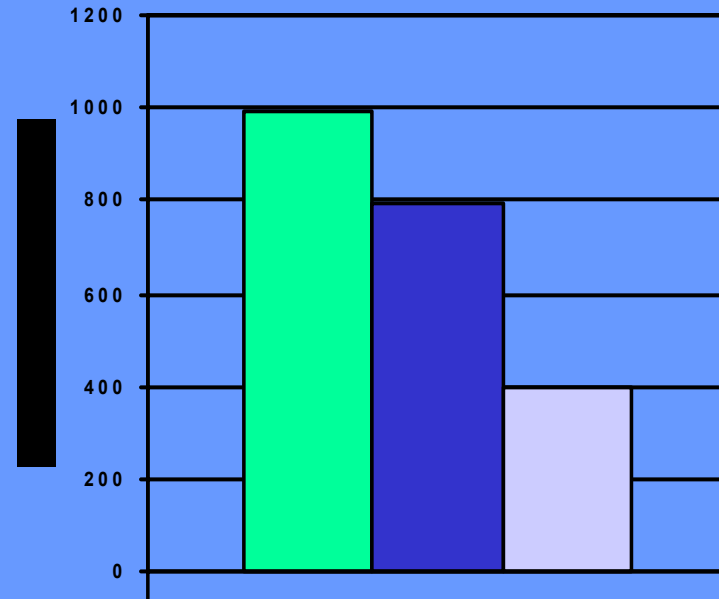
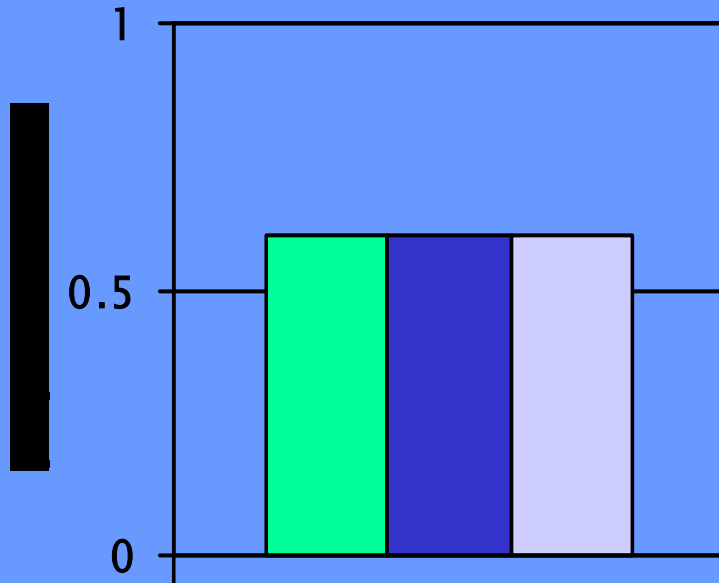




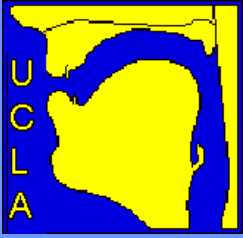
# EGG measures: males

CQ

Velocity symmetry

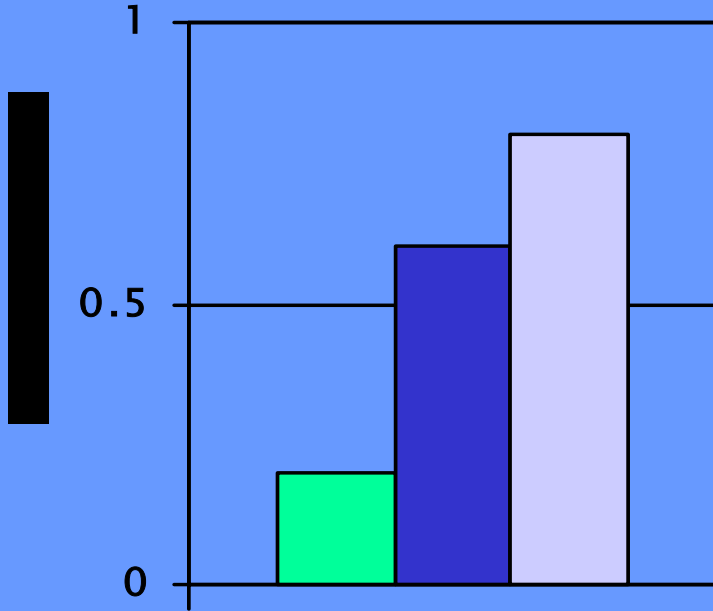


- Velocity symmetry (not CQ) distinguishes the 3 phonation categories for the male speakers
  - Suggesting that the male speakers' phonations arise from differences in closing abruptness

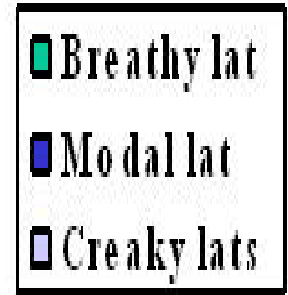
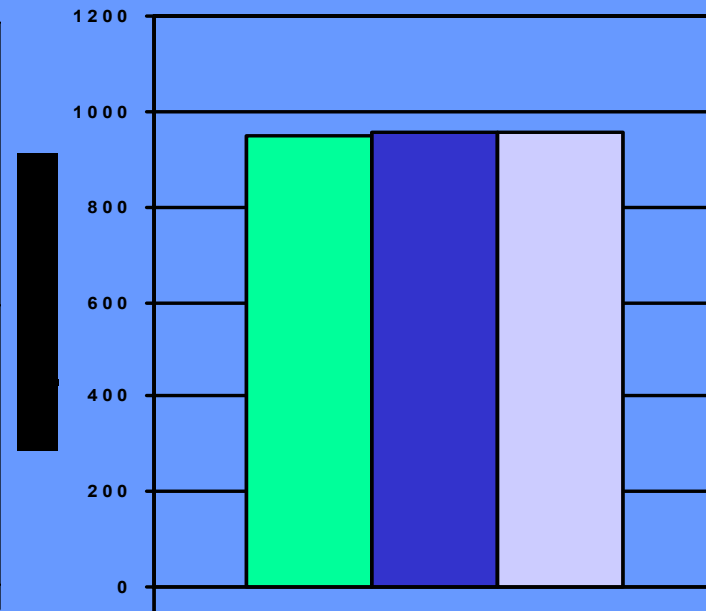


# EGG measures: females

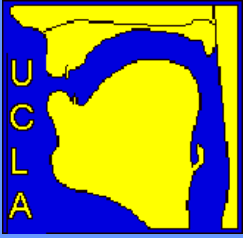
CQ



Velocity symmetry

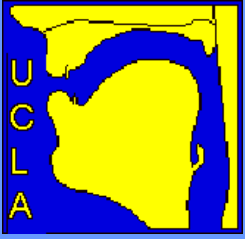


- CQ (not Velocity symmetry) distinguishes the 3 phonation categories for the female speakers
  - Suggesting that the female speakers' phonations are produced by the proportion of time the vocal folds are open during each glottal cycle



# Summary, Zapotec

Speakers	Successful measures of phonation	Suggested manner of phonation production
Male	H1-A3, Max-Min Velocity	abruptness of vocal fold closure
Female	H1-H2, Closing Quotient	proportion of cycle the vocal folds are open



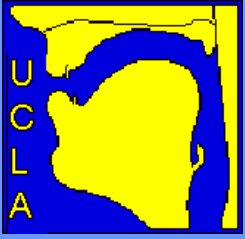
# From a continuum to a multidimensional space

- Phonation categories can be made in multiple ways
- How independent are different dimensions in a given language?
- How important is each dimension?
- Perception tests as a way to answer



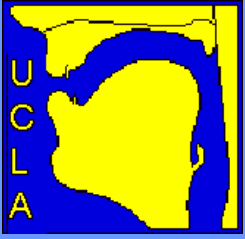
# Perception of modal vs. breathy (Esposito 2006)

- 2 experiments using different tasks and contrasting vowels from different languages
- 3 listener groups
  - 12 Gujarati (with contrast)
  - 18 American English (no contrast, but allophonic breathiness)
  - 18 Mexican Spanish (no contrast)



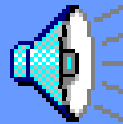
# Experiment 1: Classification

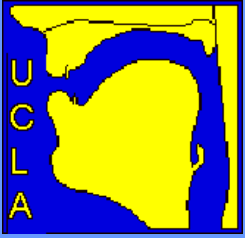
- 2 breathy and 2 modal tokens from each of 10 languages, NOT Gujarati
- Male speakers, /a/ vowels after coronals
- Discriminant analysis of this sample identifies **CPP, H1-H2, H1-A2, and H1-A3** (in that order) as most useful in distinguishing breathy vs modal





# A mix of talkers and languages

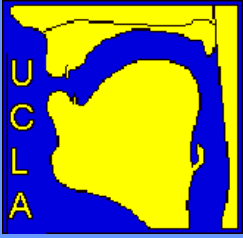
- Gujarati contrast is made simply by **H1-H2**, so the sample offers a greater variety of dimensions than Gujarati listeners are used to attending to
- Languages/talkers differ in breathiness:
  - Fouzhou (breathier) vs. Mong



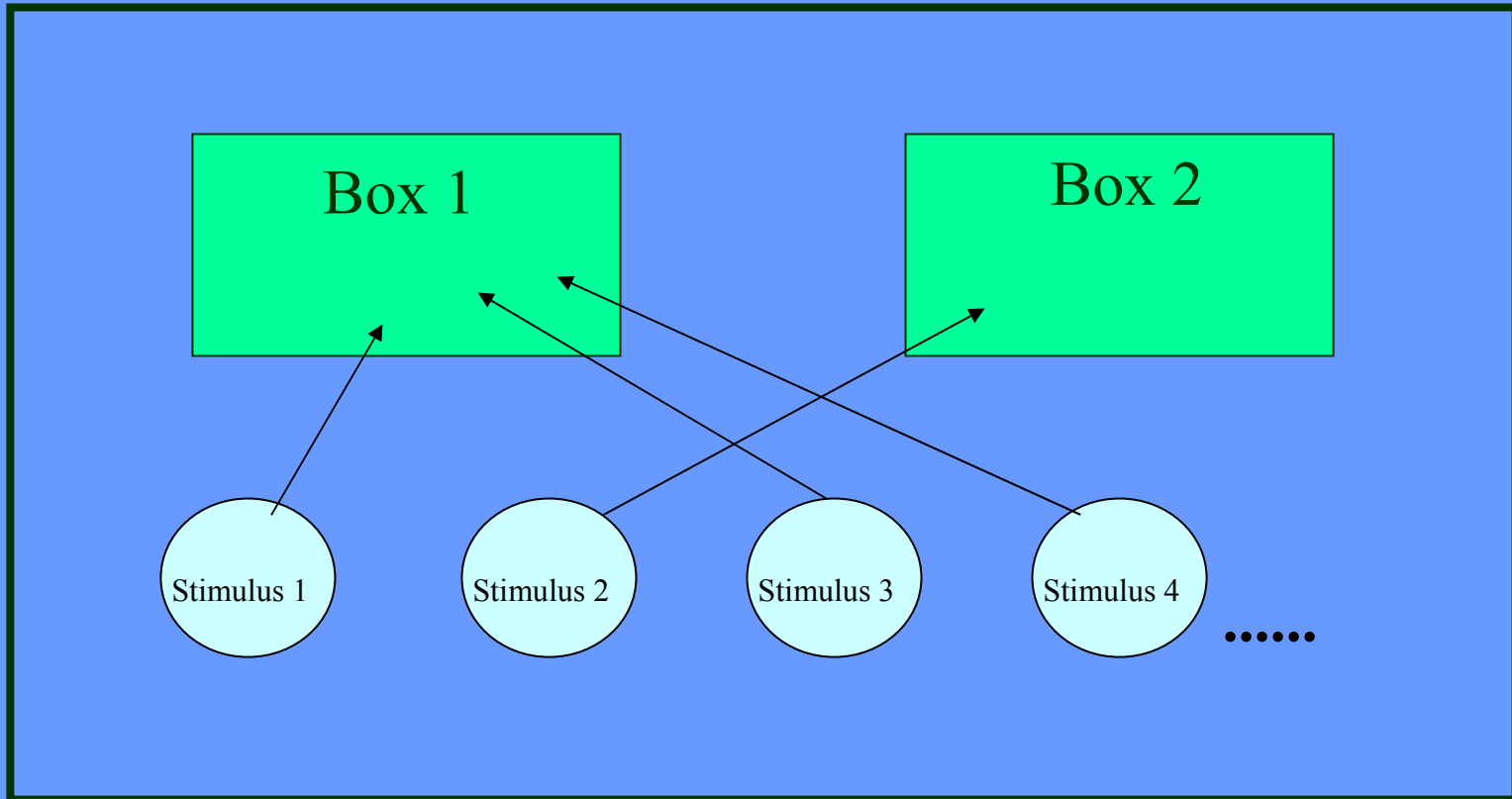


# Stimulus control

- Eliminate differences in duration and F0 by resynthesizing all tokens to 250 ms and 115–110 Hz F0
- Audio comparison of a Mazatec example:
  - Original (whole word) 
  - manipulated (vowel) 



# Visual free sort, schematic screen display

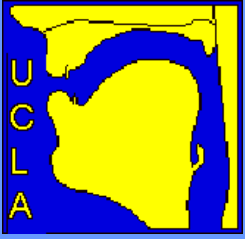


arrows represent one possible sorting of the stimuli



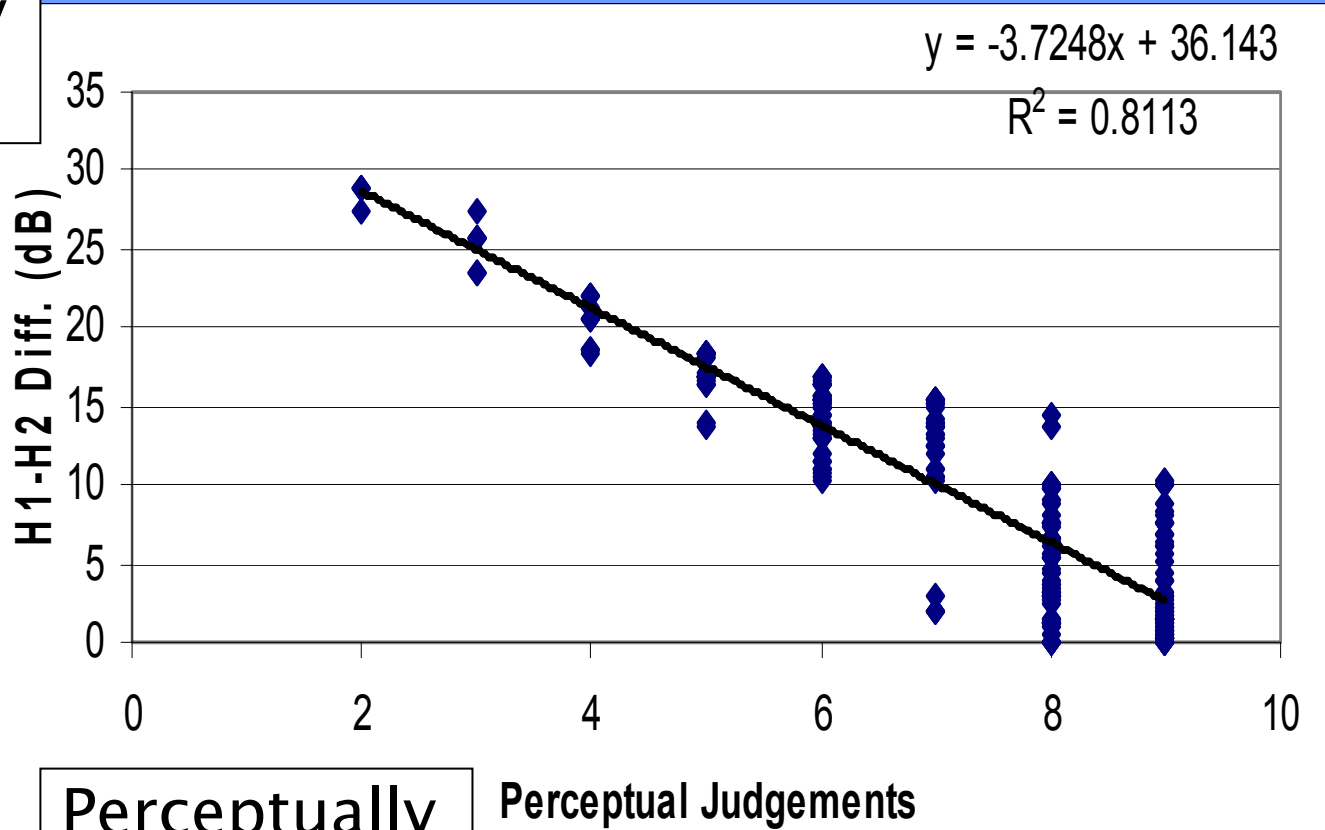
# Comparing responses across listeners

- For each pair of stimuli, how often did listeners put them in the same box, vs. in different boxes? (perceptual similarity)
- For each pair of stimuli, how different are they along each of the physical dimensions measured? (acoustic similarity)
- How are these related for each listener group? (correlations)

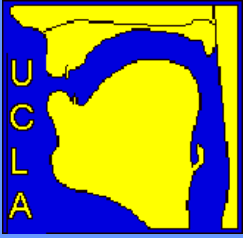


# H1-H2 and perception, Gujarati listeners

Physically different

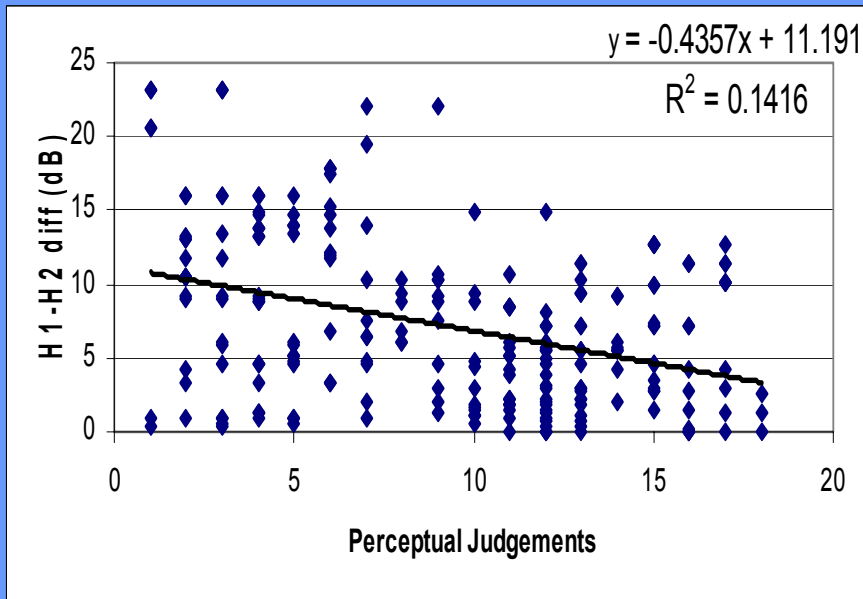


Perceptually different

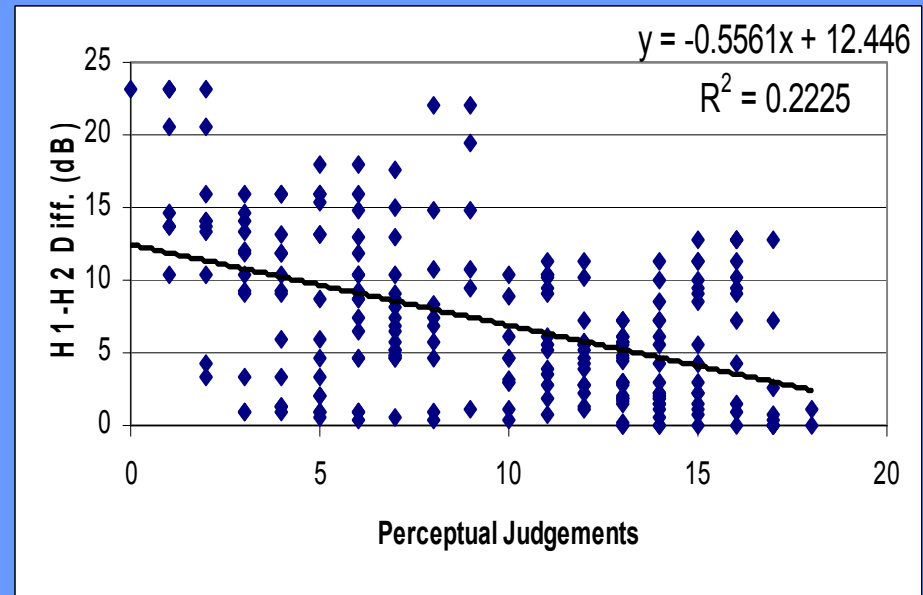


# H1-H2 and perception

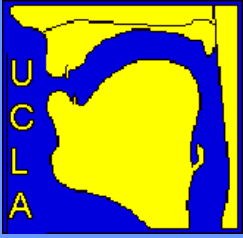
## English listeners



## Spanish listeners

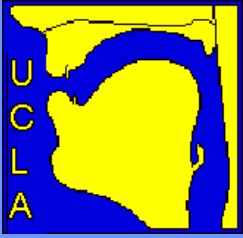


Weak correlation in both languages;  
Spanish listeners also used H1-A1



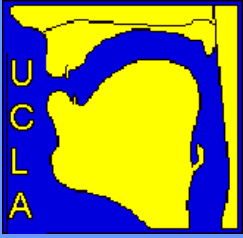
# Summary, Experiment 1

- High cross-listener consistency for Gujarati listeners
- Gujarati listeners relied only on H1–H2
- English and Spanish listeners also used H1–H2, but not consistently or well
- No listener groups used Cepstral Peak Prominence; though it was highest in the discriminant analysis, the total range of values was small



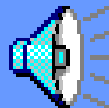
# Experiment 2

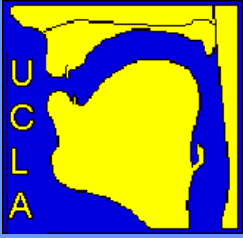
- Multidimensional perceptual spaces derived from similarity ratings of every pair of tokens in the stimulus set
- Same listeners as Experiment 1
- Different stimuli



# Experiment 2 stimuli

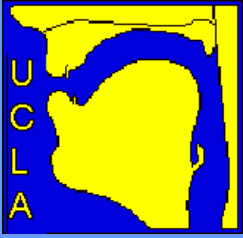
- All stimuli from Mazatec
  - 3 male speakers
  - 16 tokens from each speaker
  - Resynthesized as in Experiment 1
- Discriminant analysis identifies **H1-A2** as the best discriminator for these 40 tokens, followed by **H1-H2**
- Talkers differ in degree of breathiness





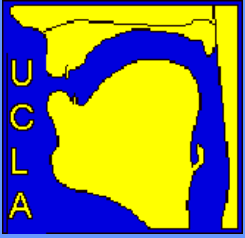
# Similarity ratings and MDS

- All possible pairs (within each talker separately) presented for similarity ratings (using an on-screen slider)
- Multidimensional scaling of ratings to derive perceptual spaces for individuals and for groups
- Perceptual dimensions related to acoustic measures by correlation

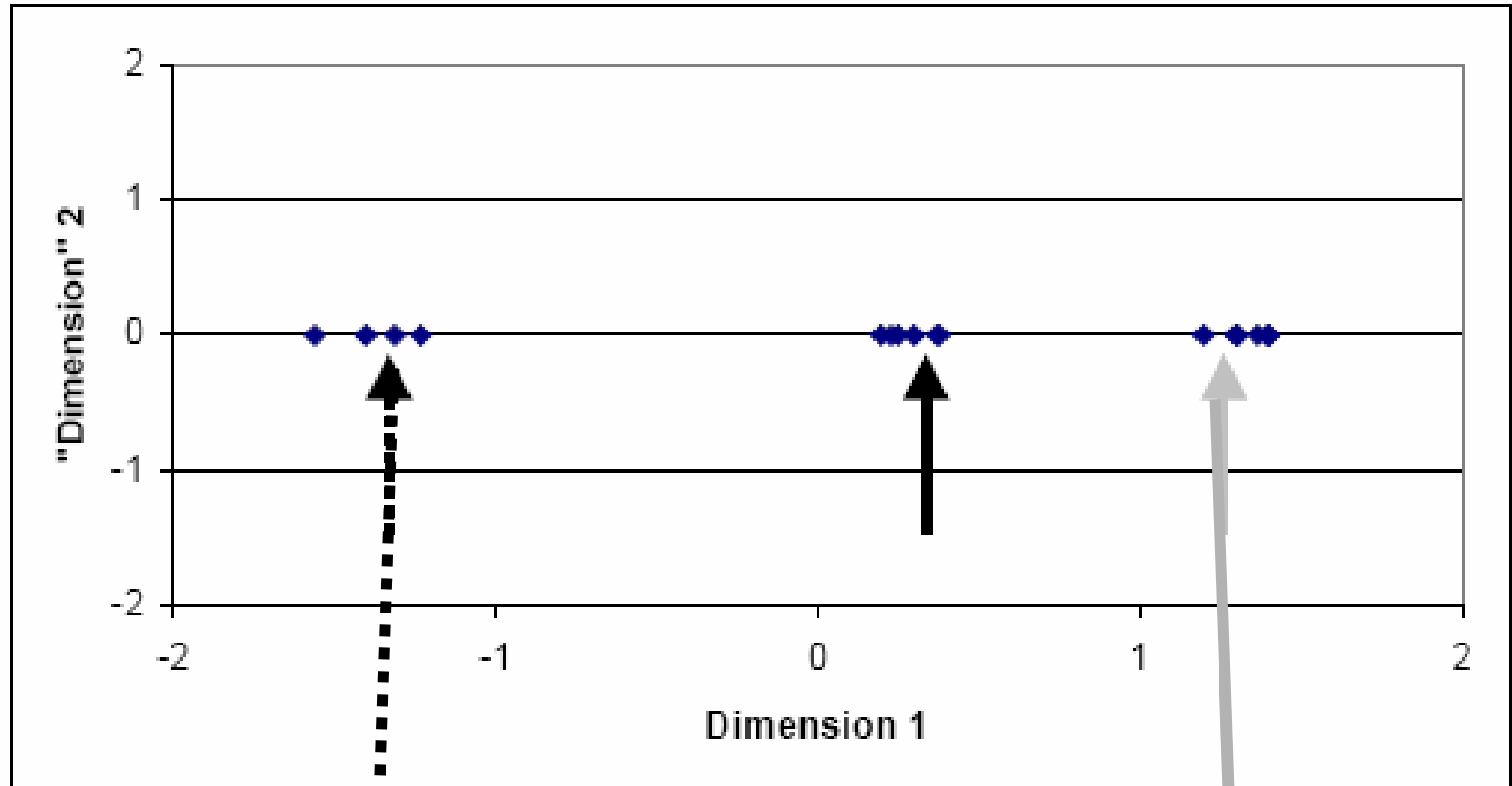


# Results

- No listeners used H1–A2 (the best one)
- English and Spanish listeners were inconsistent
- English listeners weakly used H1–H2 and CPP; Spanish listeners weakly used H1–A1 and H1–H2
- Gujarati listeners consistently relied on H1–H2, but distinguished 3 perceptual clusters rather than 2 (modal, breathy, beyond breathy)



# Sample Gujarati space with three clusters



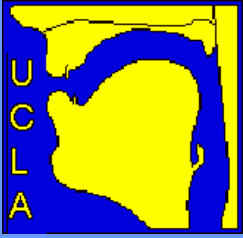
$H1-H2 \geq 15 \text{ dB}$

$H1-H2 \leq 5 \text{ dB}$



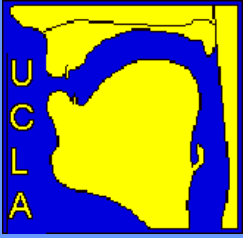
# Summary of perception experiments

- Gujarati listeners, experienced with a modal–breathy vowel contrast based on H1–H2, consistently relied on H1–H2 in perceiving vowels from several other languages
- English and Spanish listeners were inconsistent, relying weakly on a variety of (sometimes weak) correlates



# Phonation, F0, tones

- Phonation varies with tone in some tone languages
- Perhaps more general variation across languages, subtle because within modal range?
- Voice quality might be used to recognize tones
- Voice quality might be used to calibrate speaker's F0 range



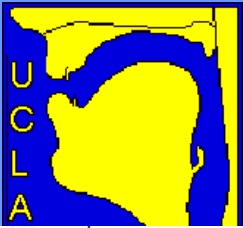
# Preliminary foray

- 4 tones of Mandarin (tones 3 and 4 known to occur with creak)
- High, Low level tones of Bura (Chadic, Nigeria)
- One male speaker of each language

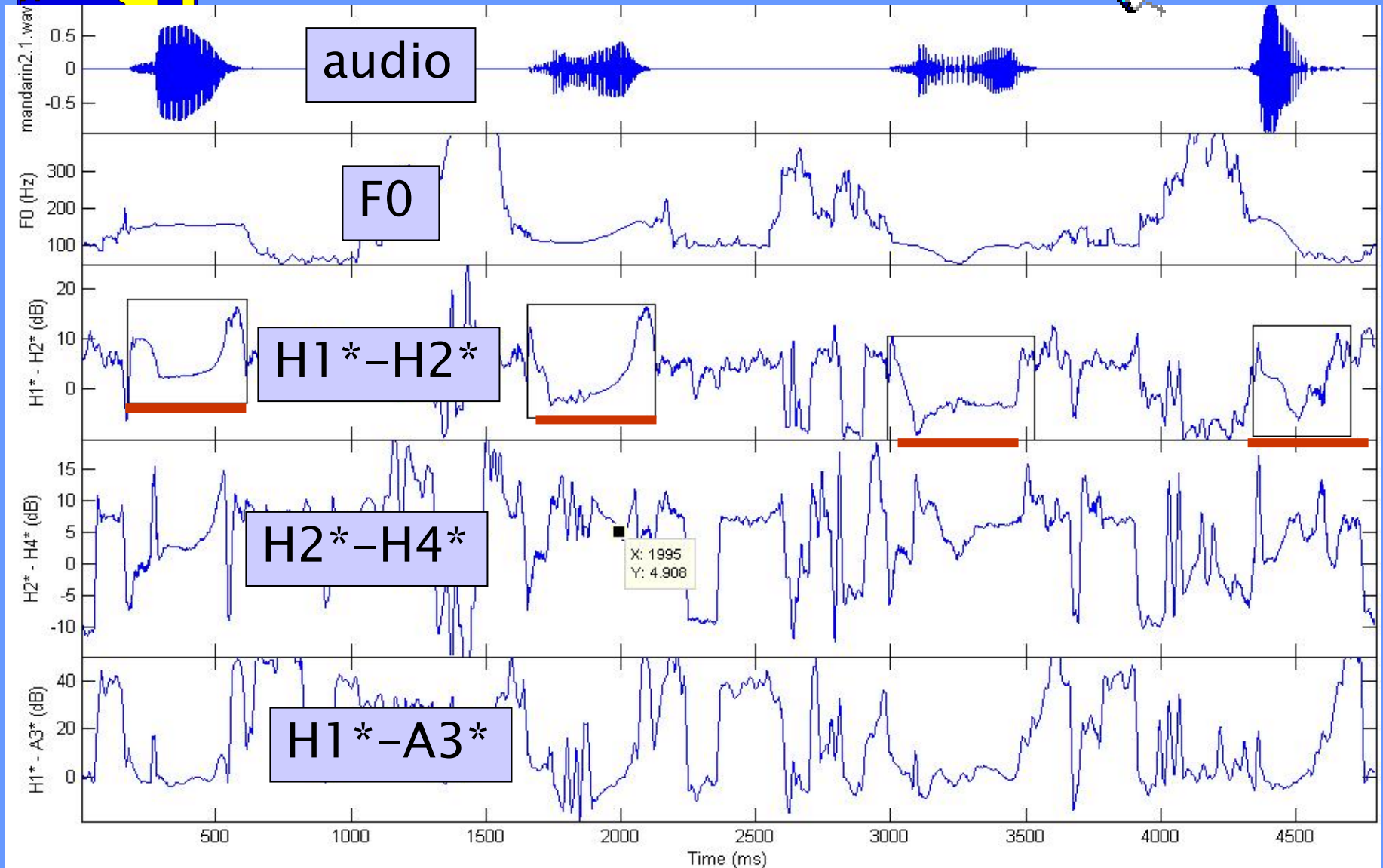
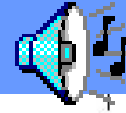


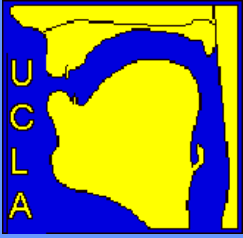
# Refining harmonic measures

- H1 and H2 are very sensitive to frequency of F1, which limits vowel comparisons
- Inverse filtering recovers the voice source, but is not always practical
- Iseli & Alwan (2004), Iseli, Shue & Alwan (2006) provide corrections for higher formant frequencies and BWs
- $H1^* - H2^*$ ,  $H2^* - H4^*$ ,  $H1^* - A3^*$



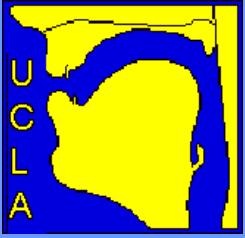
# Sample output: Mandarin





# Mandarin

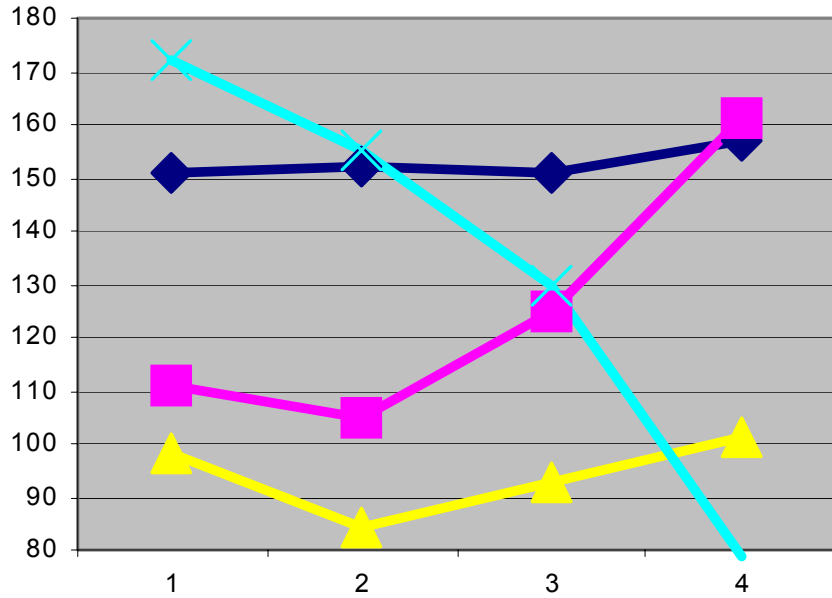
- $H1^* - H2^*$  is most related to  $F0$
- Low  $F0$  is creakier and high  $F0$  breathier
- [ Compare Iseli et al. similar result for English: below 175 Hz,  $F0$  is positively correlated with  $H1^* - H2^*$  ]
- $H1^* - H2^*$  is positive, zero, or negative with high, mid, or low  $F0$  tone onset (next slide)



# F0 and H1\*-H2\*

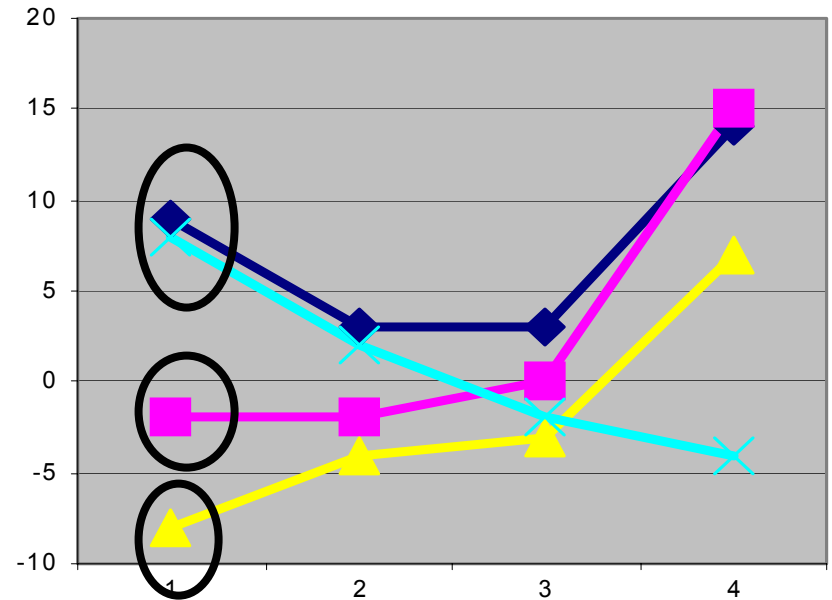
## 4 timepoints per vowel

F0 of 4 tones

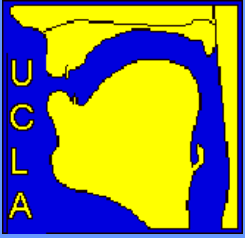


timepoints



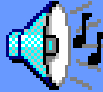
H1-H2 of 4 tones

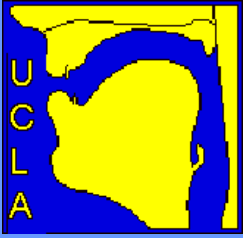


timepoints



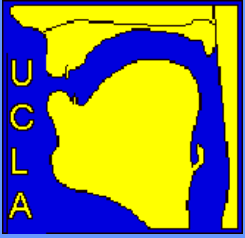
# Bura tone samples

- Larger sample, multiple tokens per tone
  - Examples of High tone 
  - Examples of Low tone 
  - Example of Low–High sequence 
- 1 measure per vowel at mid–vowel
- Compared by exploratory ANOVAs



# Bura tones and $H1^*-H2^*$

- $H1^*-H2^*$  does NOT vary with tone or with F0
- Even though speaker's F0 is in the range identified by Iseli et al.
- Different from English, and from the Mandarin sample



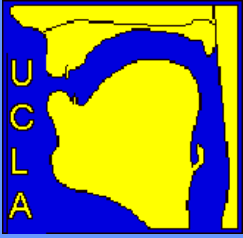
# Bura tones and other measures

- $H2^* - H4^*$ ,  $H1^* - A3^*$  are higher for Low tones, though not correlated with  $F0$
- i.e. Low tones are breathier: opposite result from Mandarin, and based on abruptness of closing rather than OQ
- Discriminant analysis with just voice measures uses  $H1^* - A3^*$  to get 57% of tokens' tones correct



# Summary, tone foray

- Mandarin and Bura tones have opposite relations of tone and voice, on different dimensions
- Within each language, voice quality could offer information to listeners about a speaker's tones



# Conclusion

Linguistic voice quality is a rich yet relatively under-studied area.

Phonation contrasts are multi-dimensional, and listeners with different language experience attend to different dimensions.

Better understanding of linguistic contrasts could help with other areas in which voice quality is important.