Poster 5aSCb6

Perceptual importance of the voice source spectrum from the second harmonic to 2kHz.

Jody Kreiman,* Marc Garellek,* and Christina Esposito+

*UCLA and +Macalester College

Introduction

- This poster examines the perceptual importance of the source spectral slope from the second harmonic (H2) to 2 kHz.
- Part of an overall effort to devise a model of voice quality comprising perceptually-valid acoustic attributes (a psychoacoustic model).
- Two sections:
  - Acoustic and perceptual analyses of the spectral slope from H2-H4 and from H4-2 kHz.
  - Data on the importance of spectral components in perception of voice quality contrasts in Hmong.

Part I: Describing the source spectrum and assessing perceptual importance of H2-H4/H4-2 kHz

- Based on previous work, we assume that perceptually-important variability in the harmonic source can be characterized by 4 acoustic parameters: the spectral slope from H1-H2, H2-H4, H4-2 kHz, and 2 kHz-5 kHz.
- Question: Do these parameters covary freely, or are there characteristic spectral shapes that tend to co-occur?
- Classical view: Spectra roll off smoothly at 12 dB/octave.

Acoustic analyses

- 49 voices
  - 18 male, 31 female
  - Randomly selected
  - Assorted vocal pathologies plus normal
  - Sustained /a/
- Source spectrum modeled using inverse filtering followed by analysis by synthesis
- 4 source spectral slope features (Fig. 1)
  - H1-H2, H2-H4, H4-2 kHz, 2 kHz-5 kHz

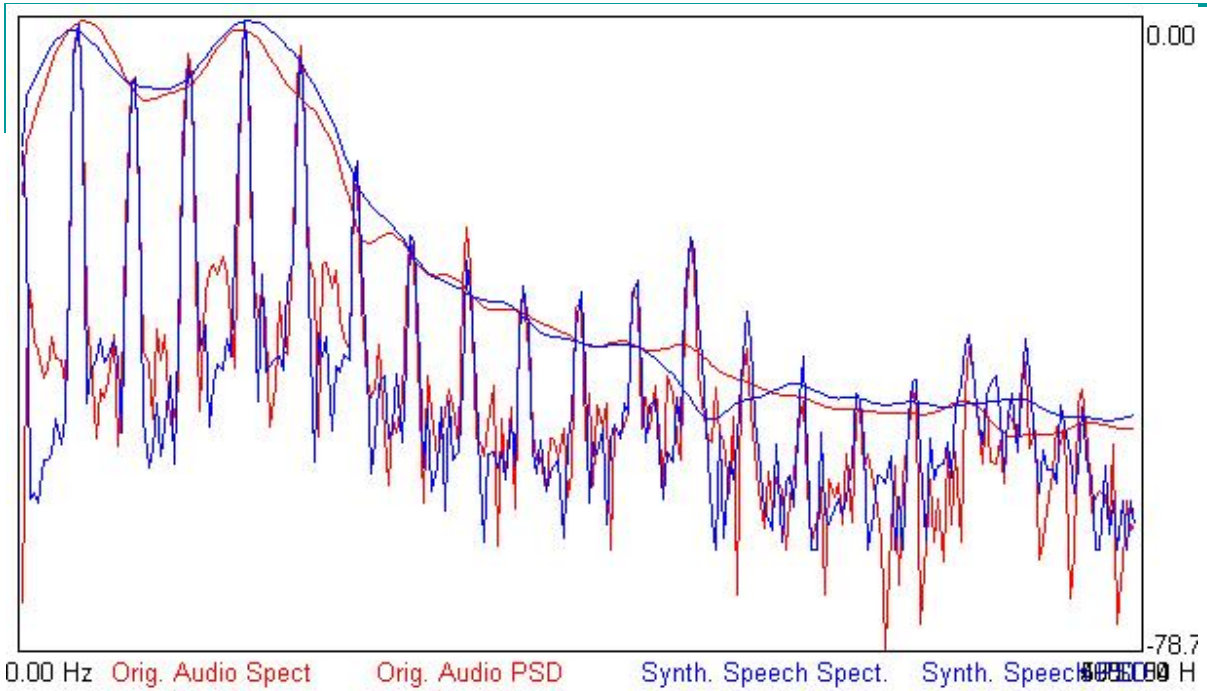
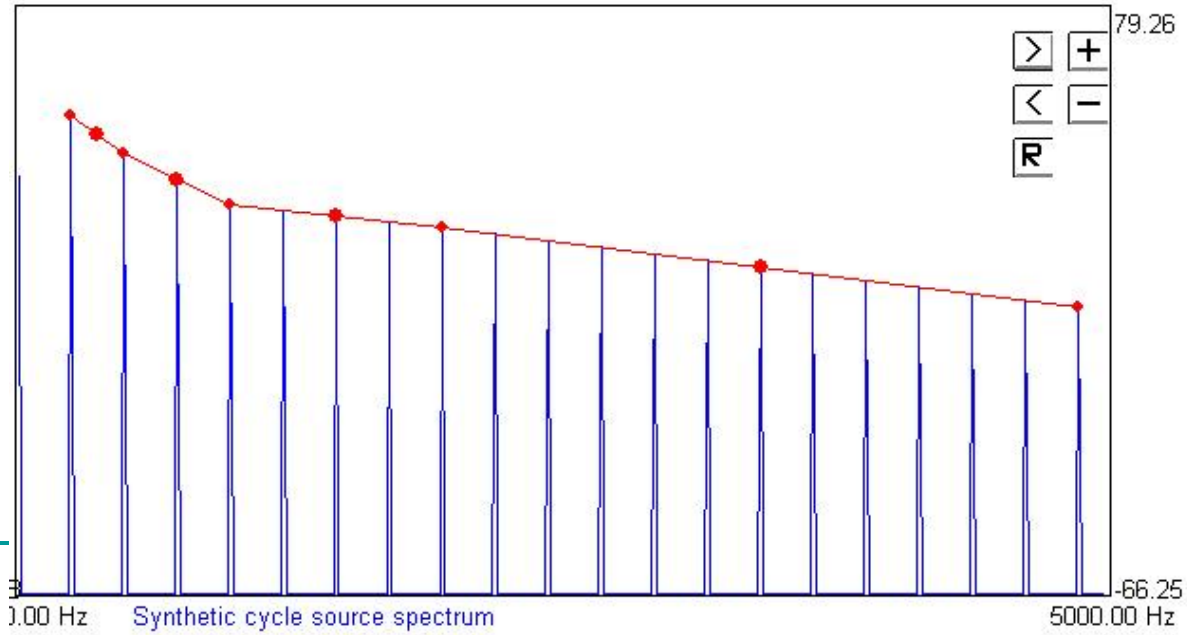


Figure 1. Natural (red) and synthetic (blue) voice spectra and the corresponding synthetic source spectrum.



Results: Patterns of spectral roll off

- Slopes of the 4 segments were uncorrelated, as expected.
- The majority of spectra were characterized by slopes that increasingly flattened (as a function of frequency) with increasing frequency* (Fig. 2a).
- Elbows occurred in this basic curve, with likelihood of occurrence and the size of the change in slope decreasing with increasing frequency (Fig. 2 b-d; $F(2, 17) = 8.48$, $p < 0.01$).
- Spectral shapes from H2-2k range from moderately concave to moderately convex.

* Same result when scaled in dB/octave or dB/Hz.

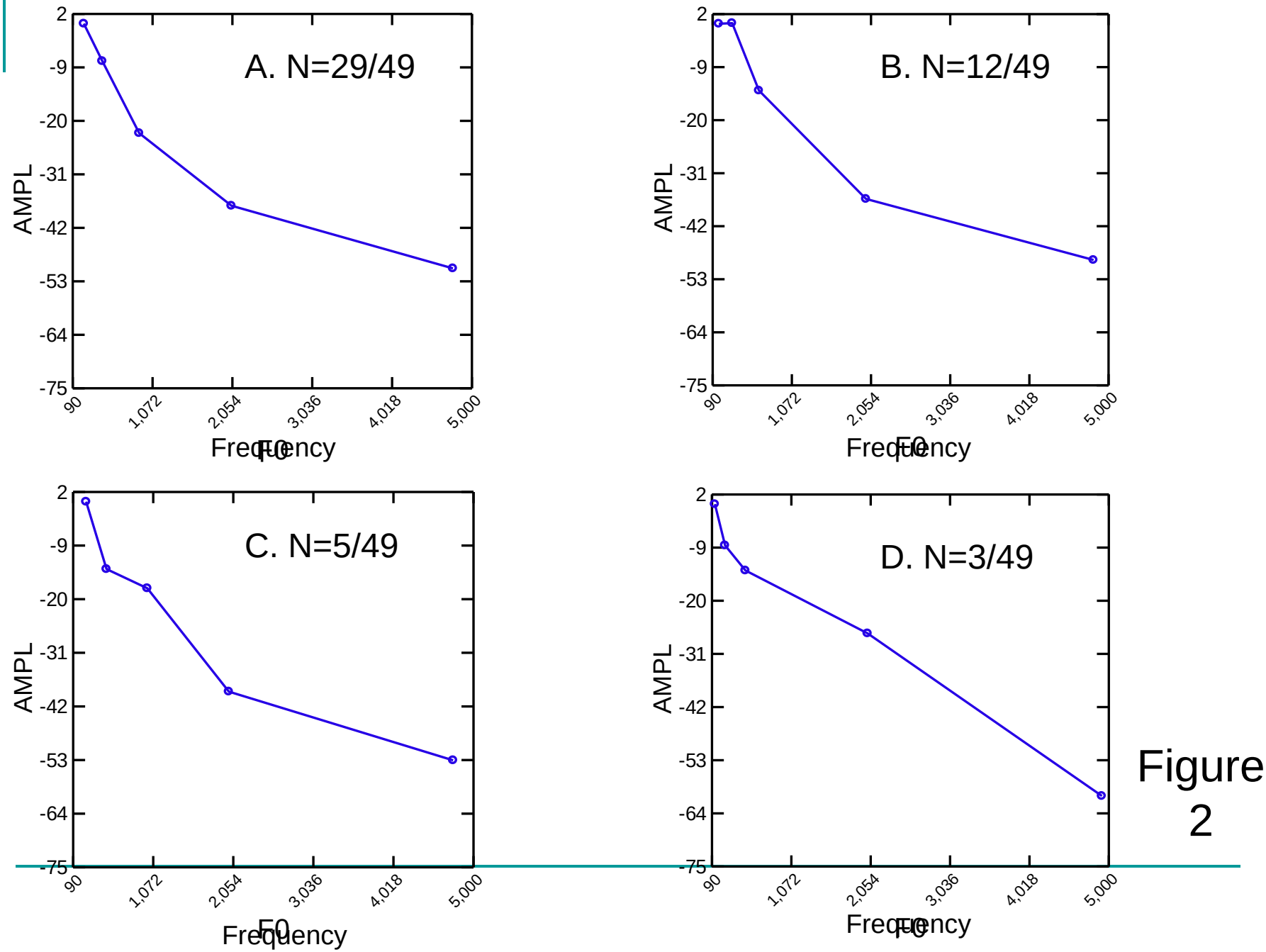
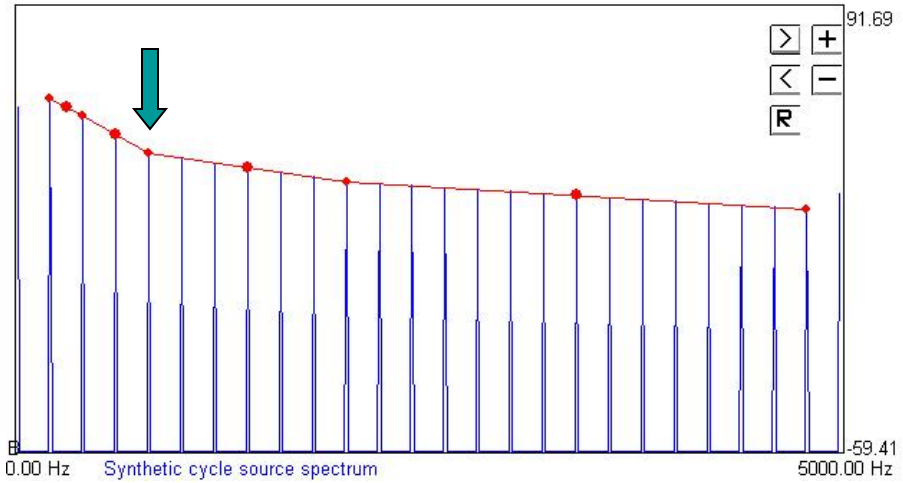

A. N=29/49
B. N=12/49
C. N=5/49
D. N=3/49
AMPL
2
-9
-20
-31
-42
-53
-64
-75
90
1,072
2,054
3,036
4,018
5,000
Frequency
F0

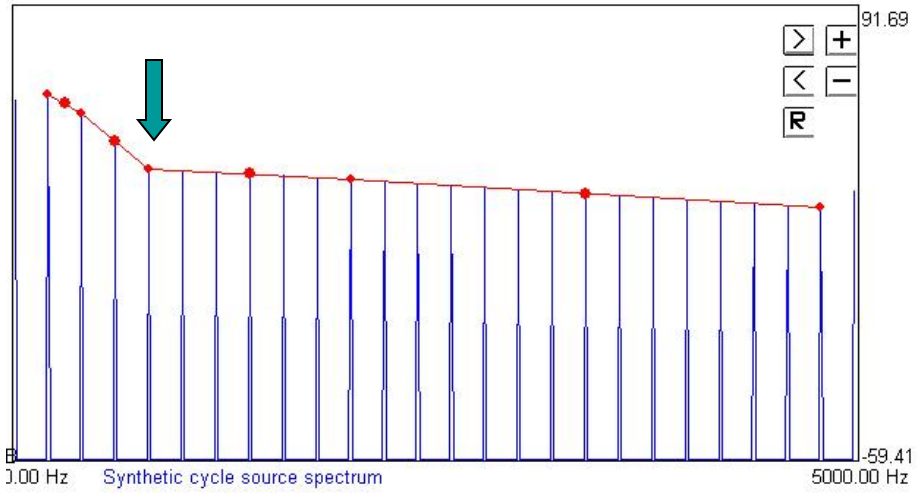
Figure 2

Preliminary perceptual evaluation

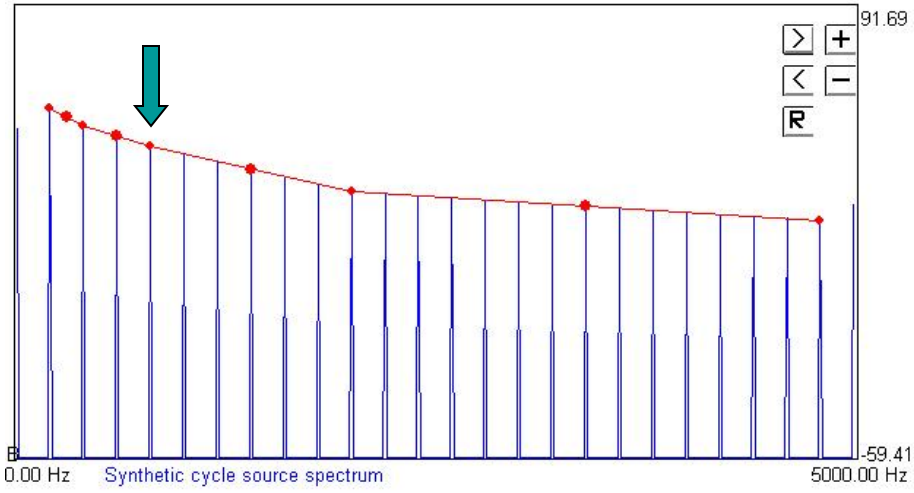
- Four voices (2 male, 2 female)
- H2-H4 and H4-2 kHz were varied in steps of 2 dB by altering the amplitude of H4 (Fig. 3).
- Other harmonic amplitudes were adjusted to maintain smooth slopes.
- H1-H2, 2 kHz-5 kHz and all other synthesizer parameters remained constant.
- One expert listener; same-different (A/X) task

Original spectrum





H4 decreased by 6 dB



H4 increased by 6 dB

Results

- JNDs for H4 amplitude (and associated spectral slopes) ranged from 2-4 dB for both increases and decreases in amplitude. Interactions with other spectral attributes remain to be investigated.
- Range of variability in H2-H4 across voices ≈ 10.4 dB; JND/range ≈ 0.29
- Range of variability in H4-2kHz across voices ≈ 18.3 dB; JND/range ≈ 0.16
- ***Consistent with the perceptual importance of these spectral features.***
- Part II investigates this possibility in more detail.

Part II: Perception of H1-H2 and H2-H4 in speakers of Hmong

- We tested the model on listeners of White Hmong.
- Hmong has two tones (–g vs. –j) that differ primarily in voice quality.
  - Cf. breathy *tag* [tâ̰] and modal *taj* [tâ]
- 15 Hmong speakers were presented with a forced choice task (breathy vs. modal word).

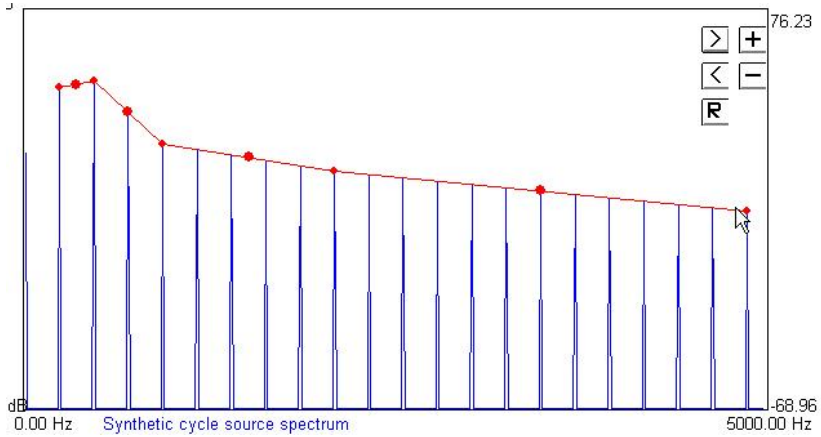
Spectral manipulations

- Synthesis modeled on a sustained /a/ vowel uttered by a healthy female speaker
  - The vowel was then shortened to a normal length for /ta/ in Hmong (about 300 ms)
  - The pitch was set to a high-falling contour starting at 280 Hz, typical for a female speaker
  - A token of onset /t/ was spliced onto the synthesized vowel.
- Controlled for factors like F0 and NHR; only spectral components were manipulated

Spectral manipulations

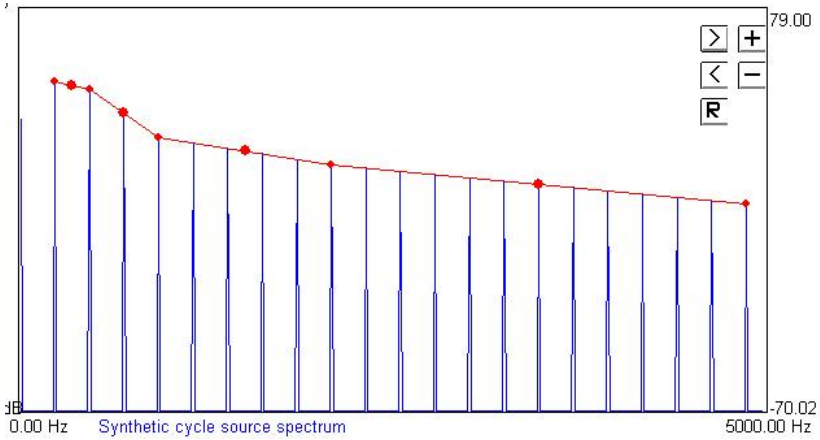
- 18 stimuli were created in five conditions:
  - Condition 1: H1-H2 varied between -2 and 15 dB (H1 manipulated; all other components held constant)
  - Condition 2: As H1-H2 increased from -2 to 15 dB, H2-H4 decreased linearly (H2 manipulated).
  - Condition 3: As H2-H4 increased from 5 to 22 dB, H4-2k decreased linearly (H4 manipulated; H1-H2 constant at 8dB)
  - Condition 4: As H4-2K increased from 10 to 27 dB, 2k-5k decreased linearly (harmonic nearest 2 kHz manipulated; H1-H2 and H2-H4 held at 8 and 22dB, respectively)
  - Condition 5: 2k-5k varied between 0 and 17 dB (highest harmonic manipulated)

Example of manipulations: Changing the amplitude of H2



← H1-H2 = -2 dB; H2-H4 = 22 dB

H1-H2 = 3 dB; H2-H4 = 17 dB →



Results

- We ran a logistic mixed-effects regression model
  - Breathy vs. modal response was the dependent variable
  - Each component was a fixed effect
  - Subject as random effect

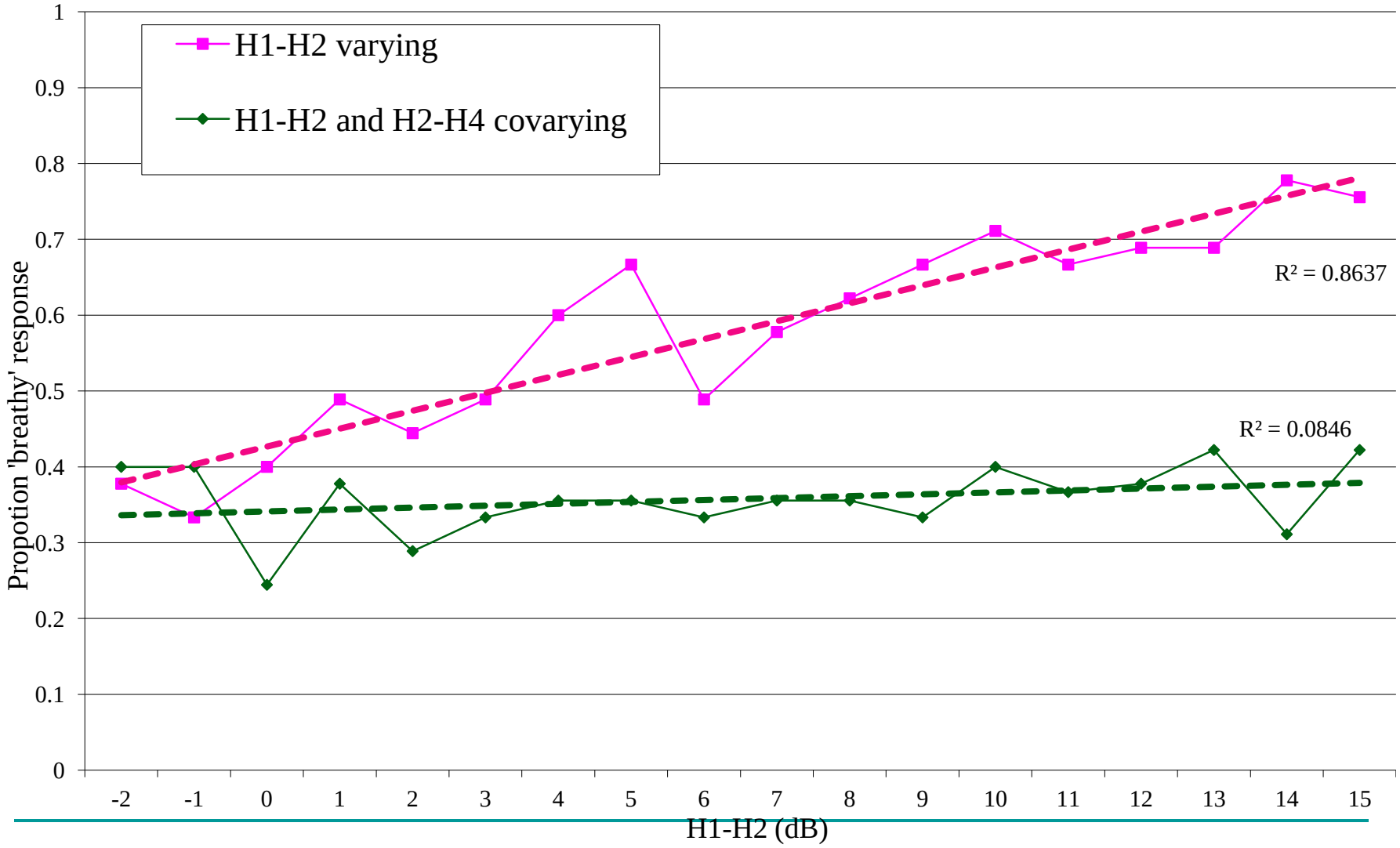
- Each model component contributed significantly to perception of phonemic breathiness, but H1-H2 and H2-H4 (independently) contribute most significantly to model fit ($z > 10$ $p<0.0001$; cf. $z = -2.2$, $p<0.05$ for H4-2 kHz, $z = -3.4$, $p<0.001$ for 2 kHz-5 kHz).

Results (continued)

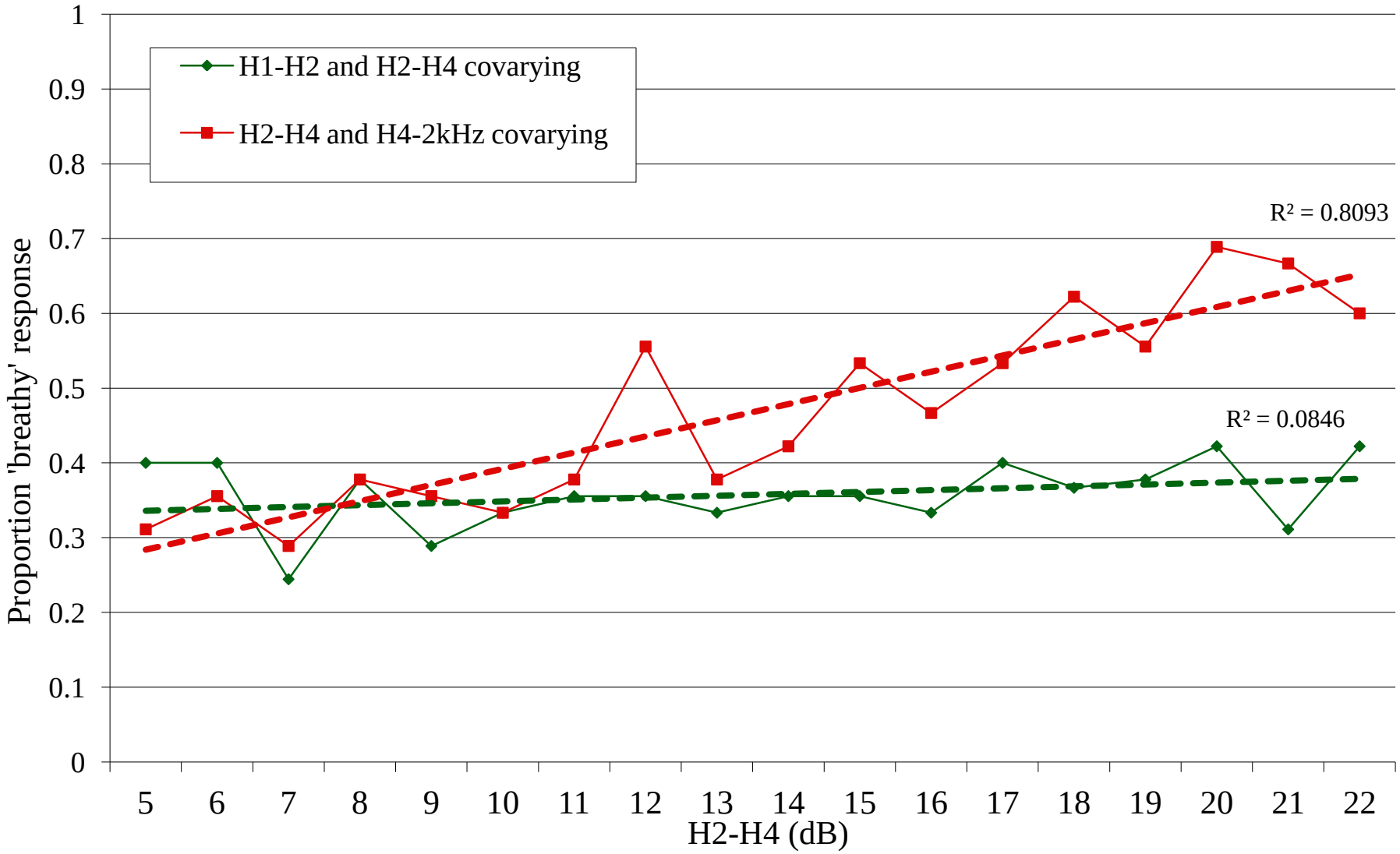
- Increase in H1-H2 and/or H2-H4 > more "breathy" responses.

- When they covary (as in Condition 2), the two components can cancel each other out:
  - High H1-H2 but low H2-H4 → modal
  - Low H1-H2 but high H2-H4 → modal

Proportion of 'breathy' responses as a function of H1-H2



Proportion 'breathy' responses as a function of H2-H4



Conclusions

- A model of the source spectrum comprising only four components (H1-H2, H2-H4, H4-2kHz, and 2kHz-5kHz) can adequately model a range of pathological and healthy voices.
- Most deviations from continuous source spectral rolloff occur in the lower frequency range (below 2kHz)
- Each of the components is perceptually useful, with H1-H2 and H2-H4 both independently responsible for determining perception of linguistically breathy voice quality.

Acknowledgments

- This research was supported by NIH grant DC01797, NSF grant IIS-1018863, and FQRSC doctoral grant 138016.