

# Perception of coarticulated speech with contrastively enhanced spectrotemporal patterns

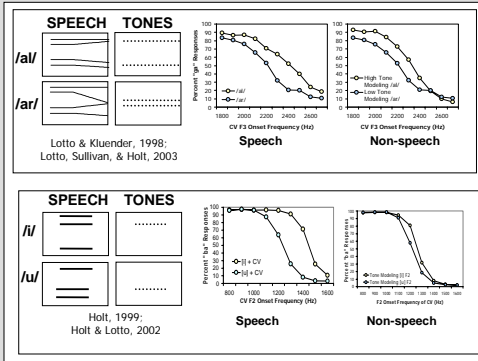
Carnegie Mellon

Lori L. Holt & Travis Wade  
 Department of Psychology & Center for the Neural Basis of Cognition  
 Carnegie Mellon University



## Introduction

Speech and non-speech exhibit similar spectrally contrastive context effects on speech perception:



→ suggests that (some) compensation for coarticulation in speech is caused by general auditory contrastive mechanisms.

**Question:** could a "dumb", general spectral contrast-enhancing mechanism actually lead to overall improvements in speech reception?

If so, what form would it have to take?

- Vowel identification improved by simple exaggeration of formant values based on surrounding values (Kuwabara, 1985), and dynamic feature information useful in speech recognition (e.g. Furui, 1986)
- Goal of present study: exaggerate spectral movements based on surroundings, observe differences in human perception
- LPC-derived log ratio area coefficients (Rabiner & Schafer, 1978) used to represent represent/manipulate spectral patterns
  - Can be effectively exaggerated to enhance spectral differences between sounds (e.g. McCandliss et al., 2002)
  - Emphasize change in coefficients over time to extend movement in the spectral domain, (roughly) exaggerating some formant trajectories, ideally removing some effects of coarticulation

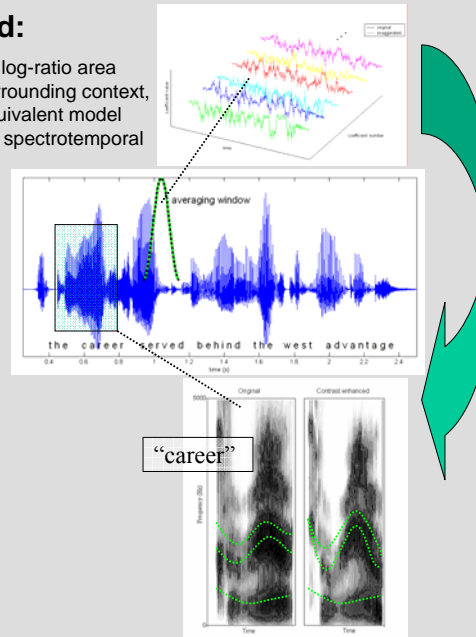
**References**  
 Furui, S. (1986) "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-34, 1, pp. 52-59.  
 Kuwabara, H. (1985) "An approach to normalization of coarticulation effects for vowels in connected speech", Journal of the Acoustical Society of America, 77, 686-94.  
 McCandliss, B.D., Fiez, J.A., Protopappa, A., Conway, M., McClelland, J. (2002) "Success and failure in teaching the [r]-[l] contrast to Japanese adults: tests of a Hebbian model of plasticity and stabilization in spoken language perception. Cognitive, Affective, & Behavioral Neuroscience, 2, 89-108.  
 Rabiner, L.R., Schafer, R.S. (1978) Digital processing of speech signals. Englewood Cliffs, NJ: Prentice-Hall.

## General Method:

Exaggerate time-varying log-ratio area coefficients based on surrounding context, enhancing vocal tract equivalent model movements & increasing spectrotemporal contrast

### Variable parameters/issues

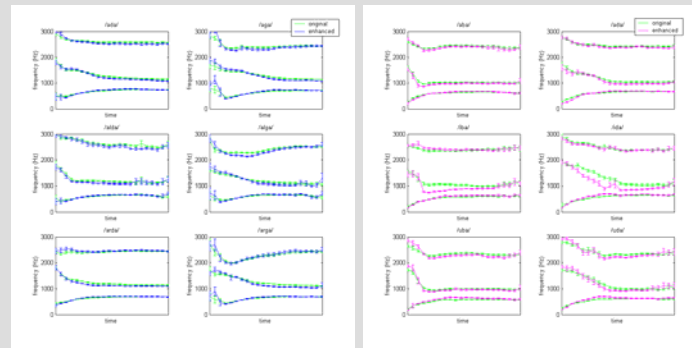
- Temporal size/ shape of integration window
- Amount of adjustment
- Treatment of following vs. preceding context
- Inclusion of intensity information
- Contrastive influences across, e.g., voicing differences



## Experiment 1: Can spectrotemporal exaggeration account for documented isolated speech/nonspeech perceptual effects?

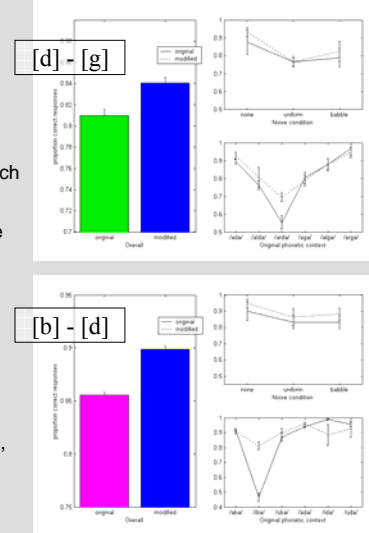
### Method

- [ba], [da], [ga] produced in different preceding vowel / liquid contexts
- [V]L]CV spectrum exaggerated automatically, based on previous context
- final CV excised, identified in isolation in 3 noise conditions



## Results

- Contrast enhancement robustly improved identification overall
- Recognition patterns resembled speech/nonspeech context effects:



- Big improvements where coarticulation obscured distinction (esp. [arda], [iba])
- Small disadvantages where context enhanced distinction (esp. [arga], [ida])
- Effects clearest in no-noise, babble contexts

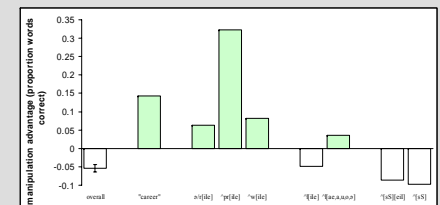
## Experiment 2: Do recognition advantages hold up for natural, connected speech?

### Method

- semantically unpredictable sentences comprised of 1-3 – syllable words produced with conversational rate, style
- listeners transcribed entire sentences in 2 noise conditions

## Results

- recognition improved for some words, particularly in instances with widely differing neighboring articulations



- disadvantages in many instances – contrast enhancement introduced distortion where compensation for coarticulation was not necessary for recognition; led to small disadvantage overall

### Conclusions

General, non-speech-specific contrast enhancement can benefit recognition of coarticulated speech in some cases, but process must be fairly "smart", taking into account at least: voicing, relative intensity, and rate