Binding Mechanisms in Speech Processing

Maria Chait*, Steven Greenberg**, Takayuki Arai***, Jonathan Z. Simon**** and David Poeppel*

**Neuroscience and Cognitive Science Program, Cognitive Neuroscience of Language Lab, Department of Linguistics, University of Maryland College Park*
*** The Speech Institute, Oakland, CA*
****Department of Electrical and Electronic Engineering, Sophia University, Tokyo*
*****Departments of Biology and Electrical & Computer Engineering, University of Maryland College Park*

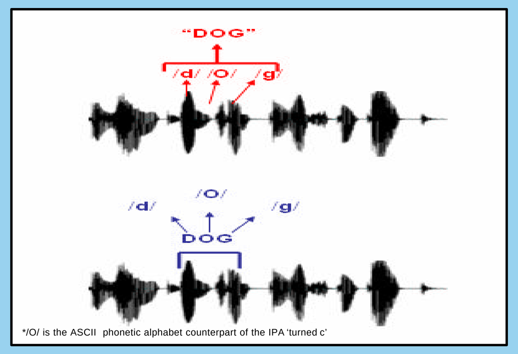
INTRODUCTION

The basic units into which the acoustic speech signal is perceptually segmented is an issue of central importance for speech research. A growing body of research points to the perceptual reality of both the phonetic segment and the syllable during the course of speech processing. These, often contradictory, findings have led to two kinds of popular models of speech segmentation: In one, the speech stream is initially segmented into phonemic segments which are later combined to create supra-segmental units. The second model assumes that the syllable is the basic unit of speech perception with the phone serving as a secondary unit of analysis (Figure 1).

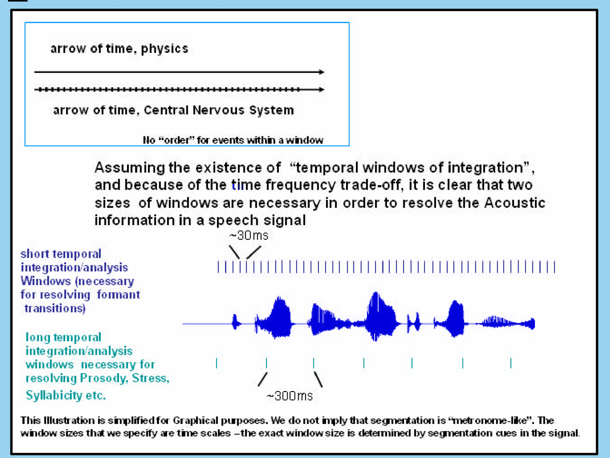
The first model accounts for fast subjects' reaction time data in phoneme-detection tasks[1] while the latter one accounts for the results obtained using the structural induction paradigm (see [2][3][4][5]). At the same time it is clear that neither model can accommodate the full spectrum of experimental results.

Here we propose a different model (Figure 3) - one that attempts to reconcile these seemingly contradictory findings and suggest a new method of systematically examining the extraction and subsequent combination of the informational constituents of the speech signal.

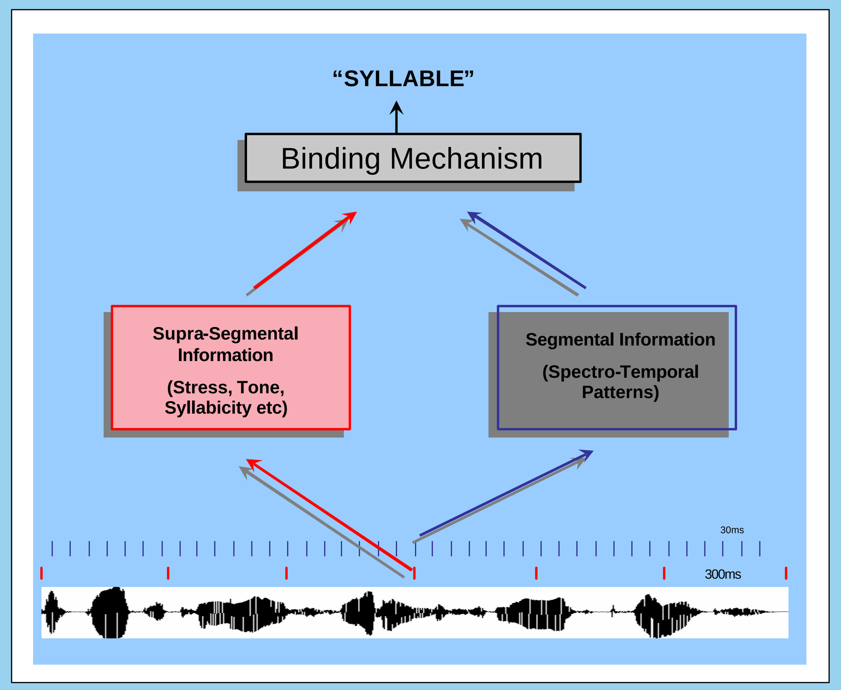
1 Existing models



2 Temporal Integration Windows



3 Suggested model



MULTI RESOLUTION ANALYSIS MODEL

The MRA model is based on the notion of **Temporal Integration Windows** (Figure 2). According to this view, the CNS treats time not as a continuous stream but as a series of temporally quantized windows and extracts data from each window[6].

We believe that segmental and supra-segmental information are extracted separately but simultaneously from the input stream from "short" (~30ms) and "long" (~300ms) windows of integration. These streams are then bound together to create a stable representation (we refer to it as the "SYLLABLE") that constitutes the input for higher order processing associated with lexical access. According to this model syllable-sized units as well as phoneme-sized units are **equally fundamental**. The precise type of information extracted from these temporal-integration windows depends on phonological and prosodic constraints specific to the listeners native language.

MODEL PREDICTIONS

1. Both low frequency (supra-segmental) and high frequency (segmental) information are necessary for successful speech processing.
2. The two streams are initially analyzed separately.
3. The binding of the two streams occurs with a delay of approximately 200-300 ms.
4. Once the stable representation has been created, we expect segmental and supra-segmental information to be perceptually integral (interference effects are symmetric)[7] [8] [9]

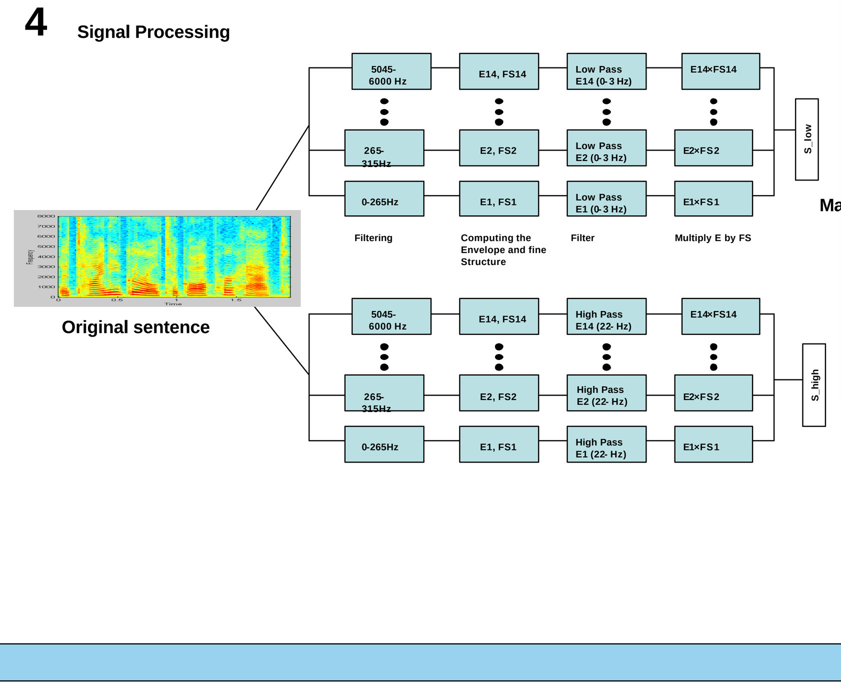
METHODS

Our signal processing technique (Figure 4) is an extension of Drullman's [10] analysis-resynthesis scheme (developed for experiments in [11]) and is based on overwhelming evidence as to the significance of the temporal envelope of the acoustic signal for successful speech processing ([12] [13] [14]).

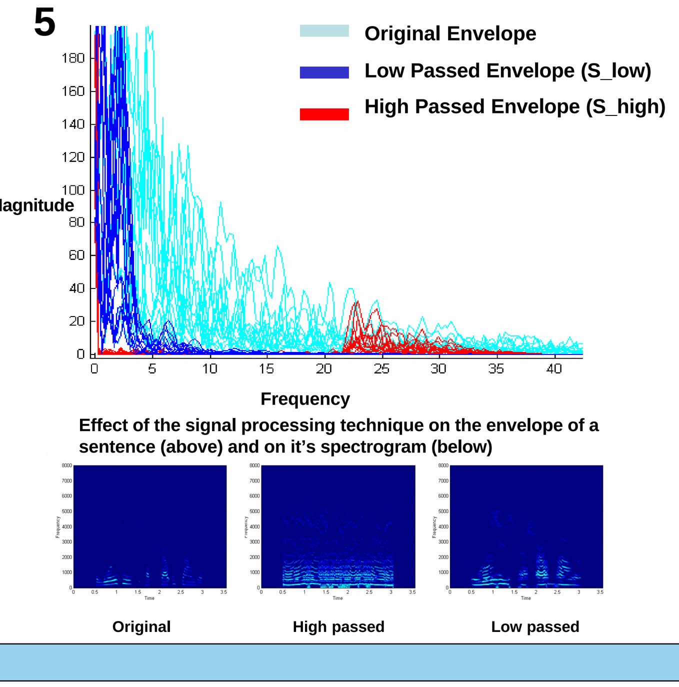
The original wide band speech signal is split into 14 frequency bands with an FIR filter bank spanning the range 0-6kHz spaced in 1/3 octave steps across the acoustic spectrum. The amplitude envelope from each band is computed by means of a Hilbert transform and then either low (0-3Hz) or high (22-40Hz) band passed before being combined again with the original carrier signal.

The result for each original signal (S) is S_low and S_high, containing only low or high modulation frequencies, which correspond to the extraction of supra segmental and segmental units (Figure 5) .

4 Signal Processing



5



Effect of the signal processing technique on the envelope of a sentence (above) and on it's spectrogram (below)

EXPERIMENT 1

Subjects: Data were obtained from 36 subjects (20 female, average age 22.5). All were native American English speakers, right handed and with no known hearing problems.

Stimuli: 53 Sentences from the IEEE corpus.
All stimuli were processed under 3 conditions:

0-3 Hz Low Pass	(presented Diotically)
22-40 Hz Band Pass	(presented Diotically)
0-3 and 22-40 Hz	(presented Dichotically)

Procedure:
Stimuli were delivered via Sennheiser HD580 head-phones. The presentation was counter-balanced to eliminate ear effects.
Each subject heard all 53 sentences but only one condition per sentence.
A practice block of 26 sentences preceded the experiment.

Task: subjects asked to write down what they heard as precisely as possible.

Responses were scored by counting the number of correct syllables in key words.

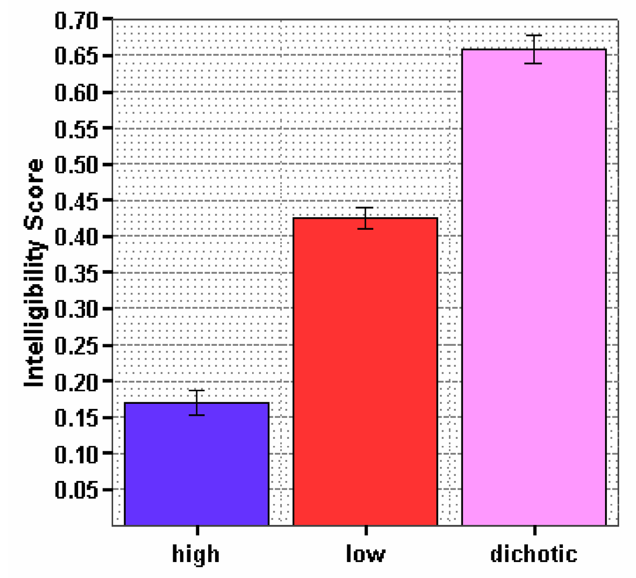
RESULTS

(1) Grand Average

Mean (S.E)

The values on the y axis reflect intelligibility scores (max=1)

All differences are highly significant at p<0.001



RESULTS

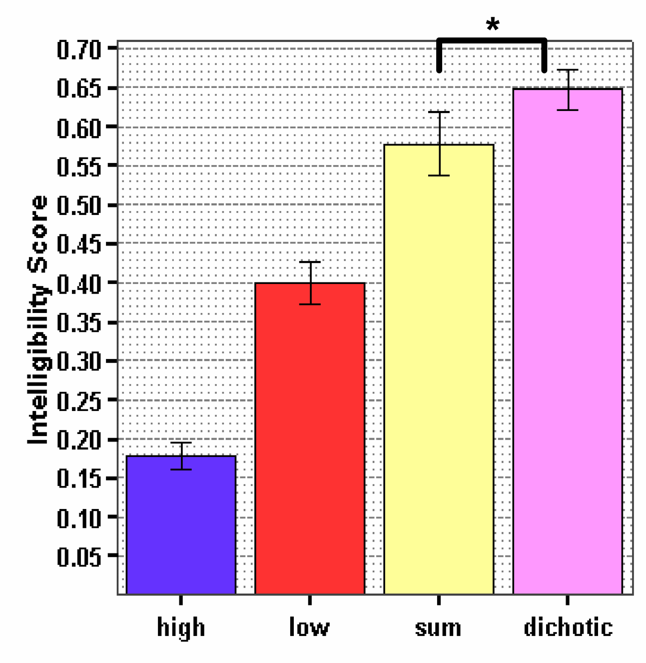
(2) Analysis by items:

The derived variable 'sum' ('sum'='low'+'high') is compared to 'dichotic' in a repeated measures ANOVA. The comparison reveals a significant difference:

$F(1,46)=5.915$, $p<0.019$ (corrected).*

This finding suggests a non linear interaction between the performance on 'dichotic' and the performance on 'high' and 'low.

* This difference is also significant in the grand average analysis



SUMMARY AND CONCLUSIONS

- Performance on the 'low' condition alone is relatively high, consistent with previous findings[10].
- Speech intelligibility is significantly increased when both low and high frequency modulation information is available to the listener. This finding is inconsistent with claims that only low frequency modulations contribute to speech understanding.
- Performance on the 'dichotic' condition is significantly higher than the sum of the performances on 'high' and 'low' conditions. This finding suggests the existence of a binding process in which the two information streams are joined together to create a composite representation that is more than the sum of its parts.
- Evidence supports the MRA model.

FUTURE WORK:

- In Experiment 2 we investigate performance on nonsense syllables (CUNY corpus).
- In Experiment 3 we introduce a time delay between the onset of S_low relative to S_high in order to investigate the temporal parameters associated with the binding mechanism.
- In Experiment 4 we look for electrophysiological (MEG) correlates of multi-time-resolution binding.

REFERENCES

[1] **Dupoux, E**. (1993) The time course of prelexical processing: The syllabic hypothesis revisited. G. Altmann & R. Shillcock (Eds.) *Cognitive Models of Speech Processing* . Hillsdale, NJ: Erlbaum. 81-114

[2] **Pallier, C**., **Sebastian-Galles, N**., **Felguera, T**., **Christophe, A**., and **Mehler, J.(**1993). "Attentional Allocation within the Syllabic Structure of Spoken Words". *Journal of Memory and Language*, 32 373-389

[3] **Finney, S.A**, **Protopapas, A**. and **Eimas, P**. (1996) "Attentional allocation to syllables in American English" *Journal of Memory and Language* 35 893-909

[4] **Pitt, M.A**., **Smith, K**. and **Klein, J**. (1998) "Syllabic effects in word processing: Evidence from the structural induction paradigm" *Journal of Experimental Psychology: Human Perception and Performance*. 24(6) 1596-1611

[5] **Yoneyama, K**., & **Pitt, M.A** (1999). Prelexical representations in Japanese: Evidence from the structural induction paradigm. Poster presented at the 14th International Congress of Phonetic Sciences, San Francisco, CA."

[6] **Poeppel, D**.(in press) "The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric' sampling in time" *Speech Communication*

[7] **Pallier, C**., **Cutler, A**., & **Sebastian, N**. (1997). Prosodic structure and phonetic processing: A cross-linguistic study *Proceedings of Eurospeech '97, 5th European Conference on Speech Communication and Technology*, Vol. 4 2131-2134

[8] **Lee, L**. and **Nusbaum H.C**. (1993) "Processing interactions between segmental and supra-segmental information in native speakers of English and Mandarin Chinese" *Perception and Psychophysics* 53(2) 157-165

[9] **Repp, B.H**. and **Lin H.B**. (1990) "Integration of segmental and tonal information in speech perception: a cross linguistic study" *Journal of Phonetics* 18 481-495

[10] **Drullman, R**., **Festen, J. M**. and **Plomp, R**.(1994a) "Effect of temporal envelope smearing on speech reception". *Journal of the Acoustic Society of America* 95(2) 1053-1064

[11] **Silipo, R**., **Greenberg, S**. and **Arai, T**. (1999) Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations, Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-99), pp. 2687-2690

[12] **Shannon, R.V**., **Zeng, F**., **Kamath, V**., **Wygonski, J**. and **Ekelid, M**.(1995) "Speech recognition with primarily temporal cues". *Science* 270. 303-304

[13] **Greenberg, S**. and **Arai, T.** (2001) The relation between speech intelligibility and the complex modulation spectrum. Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001), pp. 473-476.

[14] **Ahissar, E**., **Nagarajan, S**., **Ahissar, M**., **Protopapas, A**., **Mahncke, H**., and **Merznich, M**. (2001). "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex". *Proceeding of the National Academy of Science.* 98(23) 13367-13372

AKNOWLEDGEMENTS

This work is supported by NIH grant number DC 05660 to DP.