NON-LINGUISTIC SENTENCE-LENGTH PRECURSORS AFFECT SPEECH PERCEPTION: IMPLICATIONS FOR SPEAKER AND RATE NORMALIZATION

Lori Holt & Travis Wade

Dept. of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA
lholt@andrew.cmu.edu

ABSTRACT

Speech contexts can influence phonetic perception considerably, even across extended temporal windows. For example, manipulating spectral or temporal characteristics of precursor sentences leads to dramatic changes in categorization of subsequent vowels and consonants (e.g., Ladefoged & Broadbent, 1957; Summerfield, 1981). These findings often have been discussed in terms of speaker and rate normalization. The present study aimed to uncover precisely which types of information in the speech signal subserve such shifts in speech categorization. A series of experiments examined the influence of sentence-length non-speech precursors--series of brief pure tones--on the perception of speech segments with which they shared critical spectral and temporal properties. Across multiple experimental manipulations, the non-speech precursors affected the perceived place (alveolar, velar) and manner (stop, glide) of articulation of synthesized English consonants. Effects were observed even when non-speech precursor series were temporally-nonadjacent to the speech categorization targets and even when multiple interrupting acoustic events separated precursor and target. Both category boundary shifts and changes in graded internal category structure were observed. These results indicate that the auditory system is sensitive to both spectral and temporal information conveyed by non-linguistic sounds across sentence-length temporal windows. Moreover, this sensitivity influences speech categorization, highlighting that general auditory processing may play a role in the speech categorization shifts described as rate and speaker normalization.

INTRODUCTION

Spectral and temporal properties of speech sounds are perceived in a thoroughly context-dependent manner, at both local and more protracted levels of the speech signal. For example, immediate phonetic context influences categorization of sounds based on formant locations (Mann, 1980), as do longer-term, speaker-specific spectral patterns (Ladefoged and Broadbent, 1957). Likewise, categorization of sounds based on durational information is affected by both local and longer-term speaking rate (Miller and Liberman, 1979; Summerfield, 1981).

Prevailing interpretations of these effects invoke speech-specific normalization processes that mitigate the effects of acoustic variability introduced by coarticulation and cross-talker differences. Research employing non-linguistic sounds suggests, however, that at least the more local of these effects might derive instead from general contrastive auditory mechanisms. Just as listeners are more likely to categorize ambiguous speech targets as “ga” than “da” (“ga” has greater low-frequency energy) when they are preceded by /al/ as compared to /ar/ (/al/ has greater high-frequency energy; Mann, 1980), high- and low-frequency non-linguistic sine-wave

tone stimuli with spectral characteristics of /l/ and /r/ produce the same contrastive influence on speech categorization (Holt & Lotto, 2002; Lotto et al., 2003; Lotto & Kluender, 1998). Similarly, just as listeners hear more sounds as “wa” than “ba” (“wa” has longer formant transitions) at “faster” speaking rates signaled by a shorter following vowel, the same contrastive shift is seen in the classification of frequency-modulated sine waves in longer and shorter stimuli (Diehl and Walsh, 1989). Observations of contrastive context effects across different these classes of sounds are indicative of general perceptual interactions, rather than speech-specific normalization.

The temporal adjacency of the context-providing sounds in the experiments just described allows for the possibility that these interactions arise from well-understood local interactions in neural processing such as neural adaptation (Delgutte, 1996). Thus, it is possible that higher-order linguistic processes underlie longer-term speech context effects whereas low-level auditory perceptual interactions govern the influence of local contexts upon speech categorization. A new experimental paradigm is introduced in the present studies to investigate whether the impact of non-linguistic sounds upon speech categorization is limited to local interaction in auditory processing. The paradigm further tests whether the auditory system is sensitive to the statistical structure of spectral and temporal distributions of energy over time, and whether such sensitivity may sway listeners’ perception of following speech. Observation of a context effect under these circumstances requires that linguistic and non-linguistic acoustic information interact at higher levels of auditory processing than has previously been demonstrated. In the present studies, sentence-length pure tone melodic sequences serve as both spectral and temporal perceptual contexts for following speech sounds.

EXPERIMENT 1: SPECTRAL CONTRAST

In the first series of studies, a speech series consisted of a 9-step /ga/-/da/ continuum created by incrementing F2 and F3 frequencies of natural productions. Spectral contrast-providing context stimuli were melodies consisting of 21 70-ms (+30-ms silent interval) sine-wave tones. Melodies were composed of either “high” tones sampling 2.3-3.3kHz (*M*=2.8kHz) or “low” tones from 1.3 - 2.3kHz (*M*=1.8kHz), since it was previously observed that single tones with these mean frequencies produce a spectrally contrastive context effect on speech categorization (Lotto and Kluender, 1998). For maximum variability, tone order was randomized on a trial-by-trial basis. Each melody concluded with a 70-ms, 2.3kHz standard tone. Since these melodies were distinct, ended at the same frequency, and comprised wide, overlapping frequency ranges, any influence they showed upon categorization of following speech was expected to demonstrate listeners’ sensitivity to their long-term spectral distribution and not merely to the simple acoustic characteristics of any particular segment. This possibility was tested directly in Experiment 1a, where categorization of the /ga/-/da/ continuum was measured in the context of these two precursor types. Two additional experiments investigated the extent to which any observed effects could be attributed to more central mechanisms by examining the effects of temporal non-adjacency between context and target stimuli, introducing longer silent intervals (1b) and additional intermediate standard tones (1c).

In Experiment 1a, the /ga/-/da/ continuum was preceded by the high and low tone sequences just described, and also by intermediate tone sequences (*M*=2.3kHz) with high (1300-3300) or low (1800-2800) variance patterns. Results, shown in Figure 1, indicated that speech

categorization (overall percent /ga/ responses) was significantly impacted by the melody distributions differentiating conditions ($F(3,9)=29.3$, $p<.001$), in the expected, contrastive direction. Listeners heard more /ga/ (low F3) in the presence of higher sequences.

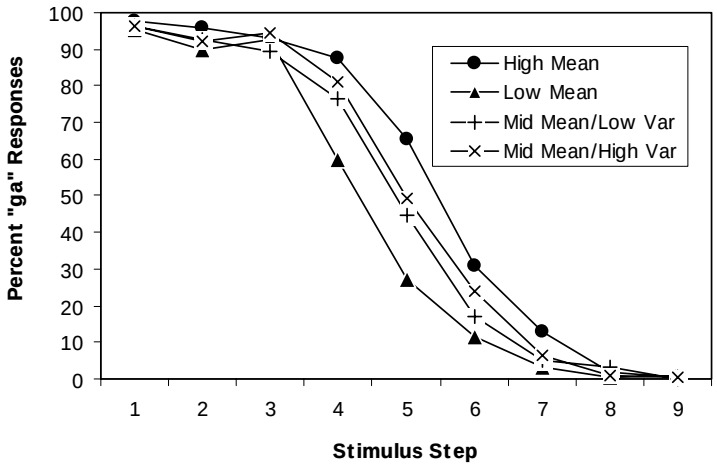

Figure 1: Categorization of /ga/-/da/ continuum based on preceding melody type

In Experiment 1b, the same speech continuum was preceded by high or low tone sequences followed by variable-length silent durations, either 100-300 ms (50-ms steps), or 500-1300 ms (100-ms steps). As shown in Figure 2, preceding melody continued to affect categorization across the ranges in both of these conditions ($p<.001$) in the same, contrastive direction.

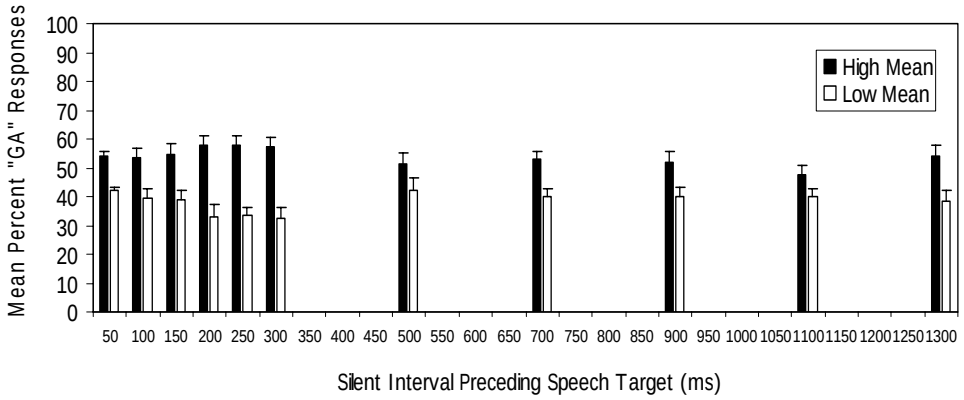

Figure 2: Percent /ga/ responses as a function of preceding melody type, silence interval (50-ms points re-plotted from Expt. 1a)

In Experiment 1c, the continuum was preceded by high or low tone sequences followed by a variable number of 2300-Hz standard tones, either 1-9 (Condition 1) or 1-13 (Condition 2). As shown in Figure 3, the preceding melody continued to affect categorization contrastively across both of these ranges ($p<.001$).

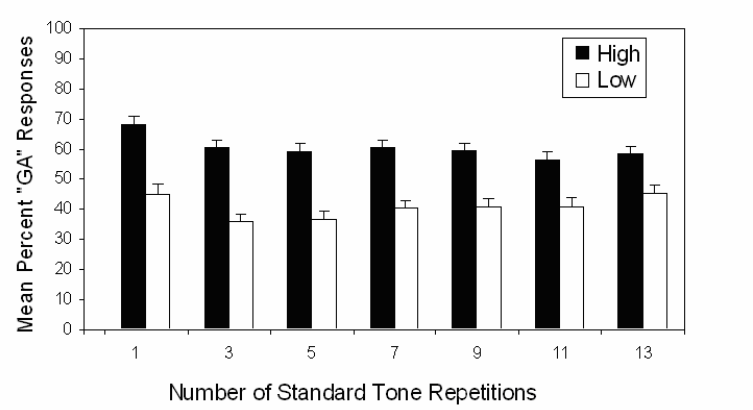

Figure 3: Percent /ga/ responses as a function of preceding melody type, number of standard tone repetitions

Thus, spectral properties of preceding tone sequences had a robust, contrastive effect on following speech stimuli, despite intervening silence (1b) and tones (1c). A similar experiment was designed to test for similar effects in the temporal dimension.

EXPERIMENT 2: DURATIONAL CONTRAST

In the second set of studies, a speech series consisted of an 11-step synthetic /ba/-/wa/ continuum created by incrementing the lengths of initial F1 and F2 transitions from 15 to 65 ms. Temporal contrast-providing context stimuli were 1.2-s melodies of either “fast” tones with durations on the order of the shortest formant transition (+40 ms initial steady state) length or “slow” tones on the order of the longest transition. Since these melodies were distinct (tone order was randomized trial-by-trial) and only onset-to-onset duration varied across types, any influence they produced upon categorization of following speech was expected to derive from long-term duration contrast rather than speech-specific rate normalization or a more local contrast effect. Three experiments were devised to ensure that this was the case. In one study (2a), melodies were either 10 110-ms (+10ms silence interval) or 30 30-ms (+10ms silence) tones, in either the F1-F2 frequency range (Condition 1; 234 – 1232 Hz) or only the F2 range (Condition 2; 769 – 1232 Hz) of the following syllable. These sequences were normalized for overall (RMS) amplitude; since different proportions (including 5-ms on-off ramps) of tone to silence dictated that maximum tone amplitude was non-constant across sequence types, an additional experiment (2b) controlled for maximum but not RMS amplitude, in sequences both with (Condition 1) and without (Condition 2) 10-ms inter-tone silence intervals. In an additional experiment (2c), absolute tone length was varied within sequences; it always averaged 110 ms (slow) or 30 ms (fast), but varied by up to 10 percent (Condition 1) or 25 percent (Condition 2) above or below these values, in order to show that subjects were sensitive to the long-term average duration and not simply the duration of a single tone adjacent to the syllable. Figure 4 shows results for these three experiments. In each case, within-subjects probit-defined /ba/-/wa/ boundary durations differed significantly ($p<0.025$) depending on precursor type, in the expected, contrastive direction. Subjects heard more /wa/ sounds (longer transitions) following fast (short tone) sequences.

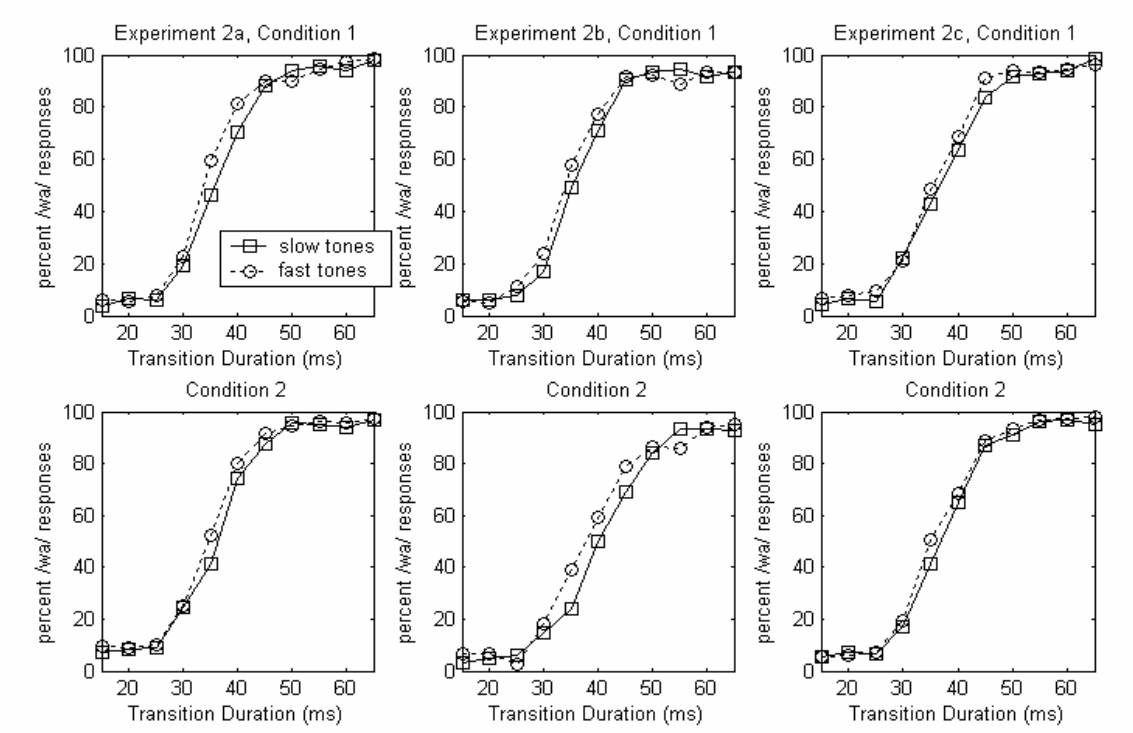

Figure 4: /ba/-/wa/ categorization patterns depending on preceding melody rate

A final experiment (2d) was conducted to determine whether preceding melody rate affects perception of the entire graded /w/ category structure (e.g. Miller *et al.*, 1997), or only the ambiguous region between /b/ and /w/. "Wa" goodness judgments were elicited for an extended /ba/-/wa/ continuum based on preceding melody rate. As shown in Figure 5, the best-exemplar range as well as the category boundary was shifted based on preceding context ($p<0.05$).

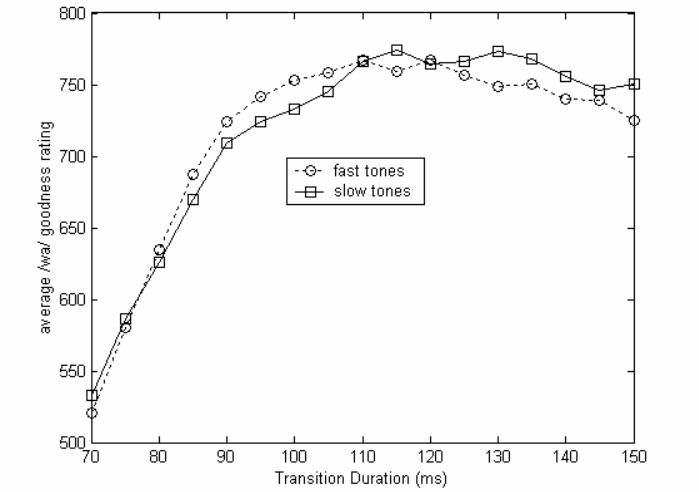

Figure 5: WA goodness judgments as a function of preceding melody

DISCUSSION/ CONCLUSIONS

The present results are informative with respect to the role of general, non-linguistic mechanisms in the context-dependent perception of speech. Categorization of speech segments based on both spectral and temporal information was influenced by context information in both dimensions provided by sentence-length non-speech precursors. Since this information was not present in the length or duration of any single precursor tone segment, but unfolded over the course of the precursor sequences, it is not likely that local, low-level contrast effects were responsible for the effects observed. Furthermore, melodic context was seen to affect both category boundaries and graded perceptual category structures, across extended intervening silence intervals and ambiguous tonal information, mirroring speaker and syllable-extrinsic rate effects seen in context-dependent speech processing. These effects suggest a greater role of general contrastive mechanisms in higher-level phonetic processing than has been previously assumed. It appears that listeners are sensitive to the statistical structure of spectral and temporal distributions of energy in auditory stimuli, in a manner that may account for these aspects of context-dependent speech perception.

REFERENCES

Delgutte, B. (1996). Auditory neural processing of speech. In W. J. Hardcastle & J. Laver (Eds.), *The Handbook of Phonetic Sciences* (pp. 505-538). Oxford: Blackwell.

Diehl, R. & Walsh, M. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustic Society of America*, *103*, 2670-6

Holt, L. L., & Lotto, A. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research, 167*, 156-169.

Ladefoged, P. and Broadbent, D. (1957). Information conveyed by vowels. *Journal of the Acoustic Society of America, 29*, 98-104.

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects of speech perception: Effect of preceding liquid on stop consonant identification. *Perception and Psychophysics., 60*, 602-619.

Lotto, A.J., Sullivan, S., and Holt, L.L. (2003). Central locus for nonspeech context effects on phonetic identification. *Journal of Acoustical Society of America, 113*, 53-56.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28*, 407-412.

Miller, J. & Liberman, A. (1979). Some effects of later-occurring information on the perception of stop consonand and semivowel. *Perception and Psychophysics*, *25*, 457-65.

Miller, J., O'Rourke, T. & Volaitis, L. (1997). Internal structure of phonetic categories: effects of speaking rate. *Phonetica*, *54*, 121-37.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 1074-95.