Rate-dependent speech processing can be speech-specific: Evidence from the disappearance of words under changes in context speech rate

Laura Dilley[1], Mark Pitt[2], Christine Szostak[3], Melissa Baese-Berk[4]

[1]Michigan State University, [2]The Ohio State University, [3]Shorter University, [4]University of Oregon
ldilley@msu.edu, pitt.2@osu.edu, cszostak@shorter.edu, mbaesebe@uoregon.edu

ABSTRACT

Function words in casual speech can be made to disappear by slowing the surrounding speech rate [1]. The current study explored the domain generality of this disappearing word effect, asking whether function words can be made to disappear by manipulating temporal properties of nonspeech context. Stimuli were short word sequences (e.g., *minor or child*) appended to precursors that were either speech, low-pass filtered speech, or tone sequences, presented at a spoken rate and a slowed rate. Across two experiments, only precursors heard as intelligible speech generated a speech-rate effect (i.e., reporting fewer function words in a slowed vs. normal-rate context), suggesting that rate-dependent speech processing can be domain-specific.

Keywords: Speech rate, spoken word recognition, lexical segmentation

1. INTRODUCTION

Are brain processes for perception of speech specialized for speech, or are they used to encode all acoustic signals regardless of their source? This question has often been studied by comparing whether a given result found with speech is also found with nonspeech. Similar findings across the two domains have been interpreted as suggestive of domain-general, as opposed to domain-specific, processes [2, 3].

One arena in which domain-specificity in speech perception has been studied is rate-dependence of phonetic perception. The identity of a consonant or vowel can be altered by changing the rate at which adjoining speech is presented [4, 5]. For example, Port [5] had listeners identify steps along a word medial /b/-/p/ continuum (*rabid* to *rapid*), with the word appended to the phrase *I'm trying to say* spoken at a fast or slow speech rate. Continuum steps which were labelled as /b/ when the speech rate was slow were more likely to be reported as /p/ when the speech rate was fast. Summerfield [6] showed that rate-dependent phoneme classification extended to other contrasts (e.g., /g/-/k/) and distinct vowel contexts (e.g., /i,ɛ/). Subsequent studies investigated which acoustic properties of the precursor contribute to the rate effect. Gordon [7] examined whether the amplitude envelope conveys speech rate (cf. [8]). In one study, he low-pass filtered a precursor (*I'm trying to say*) to make the words unintelligible while retaining low-frequency information. This manipulation shifted the boundary on a /b/-/p/ continuum. Kidd [9] showed that the boundary could also be shifted by varying precursor rhythm (e.g., slow-fast-slow-fast vs. slow-slow-slow-fast).

Although speech precursors generally produce a speech-rate effect, nonspeech precursors have yielded varied results. Gordon [7] found a shift in the /b/-/p/ boundary using a sinewave precursor whose amplitude envelope followed that of the original words; however, no rate effect was obtained when sinewaves were replaced with noise. Summerfield [6] found null results when buzz-like tones were precursors. Wade and Holt [10] found small, consistent boundary shifts using 10 long (slow sequence) or 30 short (fast sequence) sinewave tones within the F1-F2 frequency range of /b/ and /w/. Their findings, along with the results of Gordon [7], suggest that some aspects of rate-dependent speech processing are domain-general.

The purpose of the present study was to assess the generality of evidence in favor of domain-generality. We studied a novel speech rate phenomenon reported by Dilley and Pitt [1]. In casual speech, function words (e.g., *of*, *or*, *in*) can be very short. In particular, phonemes of a reduced, vowel-initial function word can be heavily coarticulated with a preceding syllable (e.g., *minor or*), causing what can be considered an elongated production of the first word (e.g., *minorrr*). (Note that there is dialectic variability in terms of which words might be expected to show heavy coarticulation with one another.) When looked at in a spectrogram, little to no acoustic information is present to demarcate the initial word boundary for the function word. In a set-up designed to elicit casual speech, Dilley and Pitt had talkers produce sentences containing sequences that are prone to blending in American English (e.g., *Anyone must be a minor or child to enter*). The speech rate of a small "target region" containing the function word (e.g., *-nor or ch-*) was manipulated, as well as the rate of

context speech surrounding this target region. The task was to type exactly what was heard. In conditions of interest here, the target region was held constant and the context speech was presented at the spoken rate or else slowed down. Slowing the rate of context speech significantly reduced listener reports of hearing a function word in the target region. The difference between spoken rate and slowed rate conditions will be referred to here as the disappearing word effect (DWE).

The present research built on results of [1] and reports two experiments which used speech and nonspeech precursors; these precursors were compared on their ability to produce a DWE. If the information in the precursor conveying rate is not unique to speech, similar results should be found across all precursor conditions. If, on the other hand, the information conveying rate is domain-specific, then only precursors that are heard as intelligible speech should yield a DWE.

2. EXPERIMENT 1

2.1. Method

2.1.1. Participants, Stimuli, and Design

Participants were 72 speakers of American English with self-reported normal hearing who received course credit for participation.

2.1.2. Stimuli and Design

Stimuli were 48 sentence fragments collected in [1]. Fragments consisted of a target region containing an ambiguous function word, and a precursor region, i.e., all words preceding the target region (e.g., *Anyone must be a minor or child*; the underlined portion is the target region). Three precursor conditions were included: clear, tone, and filtered (n = 24 per condition); furthermore, each precursor condition was crossed with two speech rate conditions: spoken rate and slowed rate. For the spoken rate condition of clear precursor stimuli, we used items collected in [1]. For the spoken rate condition of tone precursor stimuli, sequences of isochronous tones were created consisting of seven-harmonic complex tones with fundamental frequency 110 Hz. Tone duration was 100 ms with with 10 ms on/off ramps, and the number of tones matched the number of syllables in a corresponding clear speech precursor. To create the spoken rate condition of filtered precursor stimuli, precursors of stimuli in the clear precursor condition spoken-rate version were low-pass filtered at 3.5 times the mean F0 of the talker who spoke the sentence. To create all slowed rate condition stimuli, the portion of each stimulus corresponding to the precursor region in the respective precursor condition (clear, tone, or filtered) was time-expanded by a factor of 1.9, while the target region was at its spoken rate.

2.1.3. Procedure

Precursor type and speech rate were manipulated between- and within-subjects, respectively. Participants heard stimuli over headphones. After the stimulus, participants typed their response.

2.2. Results and Discussion

Responses were scored for whether the participant typed the function word in the target region. (See [1] for additional scoring details.) The proportion of function words reported in the spoken rate and slowed rate conditions is shown in Figure 1a as a function of each precursor condition. When the precursor was clear speech, a sizeable DWE was obtained. In contrast, in the other conditions, no effect of speech rate was obtained.

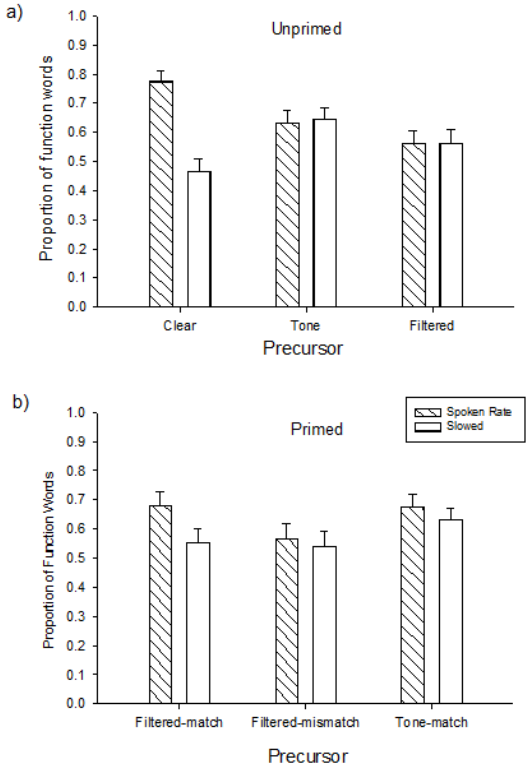

Figure 1: Proportion of function words in the target region in (a) Experiment 1 and (b) Experiment 2.

A mixed-models logistic regression analysis was performed, treating speech rate and precursor type as fixed factors and random intercepts for participants and items and random slopes for participants. Likelihood ratio tests showed the saturated model

was best-fitting; the interaction of speech rate and precursor type was reliable (b = 0.95, p < .001). Comparisons for each precursor type showed the rate manipulation was reliable only for the clear speech precursor (b = 0.30, p < .001; tone: b = -0.01, p < .95; filtered: b = 0.01, p < .95). The DWE in the clear speech condition was reliably larger than in the tone and the filtered precursor conditions (tone: b = 1.09, p < .001; filtered: b = 1.90, p < .001).

The findings of Experiment 1 thus yield evidence of domain-specificity in speech processing, since no DWE was found in the tone and filtered precursor conditions. The failure to obtain a DWE in the filtered precursor condition is surprising because low-pass filtered speech contains prosodic cues that have been shown to influence rate and speech perception in multiple ways. It may be that the time-varying cues in filtered speech (amplitude envelope, F0) are not sufficient to convey the timing information necessary to generate a DWE and that more precise temporal information at higher frequencies is needed. Of course, the presence of higher-frequency energy also makes speech intelligible, so it may be that phonetic intelligibility itself is necessary to observe the DWE. We explored this possibility in Experiment 2 using a priming experiment.

3. EXPERIMENT 2

To examine whether intelligibility is necessary to generate a DWE, we used a methodology in which a clear precursor served as the prime for the corresponding filtered precursor stimulus from Experiment 1. In the filtered-match condition, the prime was the clear (i.e., unfiltered) version of the corresponding filtered precursor. In the filtered-mismatch condition, the prime was a clear, unfiltered precursor from another phrase. If phonetic intelligibility of a context signal is necessary to convey rate information that affects speech processing, then the prime should disambiguate the precursor in the filtered-match condition, yielding a DWE. Moreover, by extension no effect of rate should be found in the filtered-mismatch condition, because the low-pass filtered precursor is unintelligible and a mismatched prime would not serve to disambiguate the words in this precursor. Note that if priming is found in the filtered-match condition, one might wonder whether any prime-precursor condition pairing that matched (on some dimension) could yield a DWE. In particular, what if a clear, intelligible speech prime were paired with a tone precursor that had been derived from it (cf. the tone condition of Experiment 1)? A DWE in this tone-match condition would be evidence of domain-generality in rate-dependent phonetic processing. If no effect of the speech rate of the prime in this tone-match condition were to be found, however, it would reinforce findings of Experiment 1 that rate-dependent phonetic processing is domain-specific.

3.1. Method

3.1.1. Participants

60 individuals from the same population as Experiment 1 participated in Experiment 2.

3.1.2. Stimuli and Design

There were three priming conditions: filtered-match, filtered-mismatch, and tone-match (n = 20 per condition); furthermore, each priming condition was crossed with two speech rate conditions: spoken rate and slowed rate. The priming and speech rate conditions constituted between- and within-subjects variables, respectively. Stimuli were created by first excising a clear (i.e., unfiltered) precursor from the respective rate condition of the clear precursor condition stimuli of Experiment 1; this clear, intelligible speech stimulus became a prime. Each prime was then prepended to one of the stimuli from the filtered precursor or tone precursor condition, as follows. To create filtered-match stimuli, a clear prime was prepended to the corresponding low-pass filtered precursor-target sequence which had been derived from it for Experiment 1. In all three priming conditions, the speech rate of the prime matched that of the precursor. To create filtered-mismatch stimuli, primes and precursor-target sequences used for filtered-match stimuli were re-paired to mismatch; that is, the prime was not an unfiltered version of the precursor to which it was prepended. The stimuli for the tone-match condition were created by prepending a clear, intelligible speech prime to the tone precursor stimulus from which it had been derived, where the tone precursor stimuli were identical to those of Experiment 1.

3.1.3. Procedure

The procedure was similar to Experiment 1. Each trial began with a prime and then the precursor-target sequence. Participants typed words spoken in the precursor-target only, and were told that the prime could help them understand the precursor.

3.2. Results and Discussion

The data were scored and analyzed as in Experiment 1. The proportion of function words is shown in Figure 1b. There is an effect of speech rate in the

filtered-match condition, but it is smaller than that in the clear-speech precursor condition in Experiment 1. The DWE did not occur in the filtered-mismatch or tone precursor conditions.

The results of mixed-models analysis with precursor condition and speech rate as fixed factors yielded a marginally reliable interaction of the variables (b = -.236, $p < .07$). Comparisons of the effect of rate within each precursor condition proved reliable only in the filtered-match condition (b = -0.127, $p < .001$; filtered-mismatch: b = -0.250, $p < .85$; tone-match: b = -0.04, $p < .60$). The DWE in the filtered-match condition was reliably larger than the filtered-mismatch condition (b = -0.703, $p < .01$) but only approached significance in the tone-matched condition (b = -0.253, $p < .11$). To evaluate the effectiveness of the matching manipulation in disambiguating the precursor in the filtered-match and filtered-mismatch conditions, we compared the accuracy of exact word transcriptions for the precursor. Mean accuracy was 71% vs. 4% in the filtered-match and filtered-mismatch conditions, respectively, confirming that priming disambiguated the precursor as intelligible speech only in the filtered-match condition.

4. GENERAL DISCUSSION

Perception of auditory events depends on accurate encoding of temporal information, something true of multiple classes of stimuli (e.g., speech, music, environmental sounds). Thus it is of interest to determine whether specialized processes may be devoted to some stimulus classes and whether there are common processes for all auditory objects.

We used the DWE to ask whether rate information conveyed by multiple types of auditory objects was equivalent by measuring the ability of each to produce a speech rate effect. The results are clear: only stimuli perceived as intelligible speech produced a DWE. In Experiment 1, function word reports across the spoken-rate and slowed-rate conditions differed only when the precursor was clear speech. In Experiment 2, a speaking rate effect was found only when the filtered precursors were disambiguated. Across multiple conditions in Experiments 1 and 2, tone precursors, whether primed or unprimed, never yielded a rate effect.

The results of this study expand our understanding of rate-dependent speech processing by identifying a condition in which the DWE depends on the precursor having specific stimulus properties. As described earlier, a phonetic boundary can be shifted by tones at different rates [10], as well as sinewave tones [7]. The results of Experiments 1 and 2 show that rate information conveyed by such precursors is insufficient to generate the DWE. Other, perhaps more detailed timing information must be present in the precursor, such as that conveyed by intelligible speech. The stimulus selectivity of the DWE suggests a degree of domain specificity in speech processing.

Wade and Holt [10] suggested that rate-dependent processing may have multiple causes, some of which could be specific to speech or language. The current data support this contention, and raise the question of how to explain them. One possibility is that domain-specific processing occurs at an abstract level of analysis. Sjerps, Mitterer, and McQueen [11] advanced this proposal in a study of contextual influences in vowel normalization. One aspect of context which varied was whether it was speech or nonspeech. Differences in responding under these two stimulus types prompted Sjerps et al. to suggest that speech-specific effects might be occurring at a more abstract level of auditory analysis than that for nonspeech contexts. The current data fit this account. The issue of which acoustic properties make the DWE more abstract, or whether these processes are unique to speech, cannot be answered here. If speech is the only natural stimulus that possesses the variation necessary to produce the DWE, then this higher level of analysis is domain-specific whether by design or not.

Other considerations suggest the rate-based phenomena require different explanations. The phonetic rate effect seems to be contrastive: the boundary shifts in a compensatory direction to ensure a stable phonetic percept. The DWE is qualitatively different, such that a change in rate causes reorganization of a stretch of speech, resulting in segments appearing or disappearing. Undoubtedly, information from e.g., semantics and syntax contribute. For a given rate, knowledge of articulator movement could be used to estimate trajectories over the target and interpret phonetic content, inferring production of one or two syllables.

Oscillator models might provide a unified account of the two rate effects. According to such approaches, oscillators tuned to temporal (quasi-) periodicities time events and coordinate actions [8, 12-14]. The DWE and phonetic boundary shifts might tap different levels of an oscillator hierarchy.

In summary, the present results suggest that the DWE may be associated with a different, speech-specific timing mechanism, in contrast to rate effects on phoneme boundaries, which appear to be largely domain-general. This finding provides new information regarding the nature of timing mechanisms for speech vs. nonspeech, suggesting at least some mechanisms relevant for understanding spoken words may be speech-specific.

5. REFERENCES

[1] Dilley, L.C., Pitt, M. (2010). Altering context speech rate can cause words to appear or disappear, *Psychological Science*, 21, 1664-1670.

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

[3] Remez, R.E., Pardo, J.S., Piorkowski, R.L., Rubin, P.E. (2001). On the bistability of sine wave analogues of speech, *Psychological Science*, 12, 24-29.

[4] Miller, J.L. (1981). Phonetic perception: evidence for context-dependent and context-independent processing, *Journal of the Acoustical Society of America*, 69,

[5] Port, R.F. (1979). The influence of tempo on stop closure duration as a cue for voicing and place, *Journal of Phonetics*, 7, 45-56.

[6] Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception, *Journal of Experimental Psychology: Human Perception and Performance*, 7, 1074-1095.

[7] Gordon, P.A. (1988). Induction of rate-dependent processing by coarse-grained aspects of speech, *Attention, Perception, & Psychophysics*, 43, 137-146.

[8] Cummins, F., Port, R.F. (1998). Rhythmic constraints on stress timing in English, *Journal of Phonetics*, 26, 145-171.

[9] Kidd, G.R. (1989). Articulatory rate-context effects in phoneme identification, *Journal of Experimental Psychology: Human Perception and Performance*, 15, 736-748.

[10] Wade, T., Holt, L.L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories, *Perception & Psychophysics*, 67, 939-950.

[11] Sjerps, M.J., Mitterer, H., McQueen, J.M. (2011). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics, *Neuropsychologia*, 49, 3831– 3846.

[12] Nam, H., Goldstein, L., Saltzman, E., Dynamical modeling of supragestural timing, in: Proceedings of the 10th Laboratory Phonology Conference, Paris, France, 2006.

[13] Port, R.F. (2003). Meter and speech, *Journal of Phonetics*, 31, 599-611.

[14] Barbosa, P. (2007). From syntax to acoustic duration: A dynamical model of speech rhythm production, *Speech Communication*, 49, 725-742.