Attention, Perception, & Psychophysics (2019) 81:1065–1075
https://doi.org/10.3758/s13414-019-01699-3

PERCEPTUAL/COGNITIVE CONSTRAINTS ON THE STRUCTURE OF SPEECH COMMUNICATION: IN HONOR OF RANDY DIEHL

A test of linguistic influences in the perceptual organization of speech

Marjorie Freggens[1] · Adam Thomas[1] · Mark A. Pitt[1]

Published online: 19 March 2019

Abstract
Research in speech perception has explored how knowledge of a language influences phonetic perception. The current study investigated whether such linguistic influences extend to the perceptual (sequential) organization of speech. Listeners heard sinewave analogs of word pairs (e.g., *loose seam*, which contains a single [s] frication but is perceived as two /s/ phonemes) cycle continuously, which causes the stimulus to split apart into foreground and background percepts. They had to identify the foreground percept when the stimuli were heard as nonspeech and then again when heard as speech. Of interest was how grouping changed across listening condition when [s] was heard as speech or as a hiss. Although the section of the signal that was identified as the foreground differed little across listening condition, a strong bias to perceive [s] as forming the onset of the foreground was observed in the speech condition (Experiment 1). This effect was reduced in Experiment 2 by increasing the stimulus repetition rate. Findings suggest that the sequential organization of speech arises from the interaction of auditory and linguistic processes, with the former constraining the latter.

Keywords Perceptual organization · Audition · Speech perception

Introduction

To make sense of the auditory world, the perceptual system must organize the incoming information into chunks that cohere over time. The comprehension of spoken language exemplifies the problem. Speech exhibits tremendous acoustic diversity, from periodicity in vowels, to aperiodicity in fricatives, to brief silent intervals when the articulators close. Yet listeners rarely experience difficulty tracking the speech of the talker, parsing it into words, and building a representation of the message, whether in a quiet or noisy environment.

The sequential organization of speech likely occurs over multiple levels of representation, being affected by both auditory and linguistic processes (Davis & Johnsrude, 2007). Very early on the perceptual system must parse the auditory scene into distinct sound sources (e.g., a voice in a noisy café) as it encodes the speech of a talker. Extensive research has explored how source perception is influenced by changes in acoustic dimensions such as timbre (vowel quality), frequency, and location (Bregman, 1990; Warren, 2008). For example, the harmonics of a vowel cohere in part because they share a common fundamental. A harmonic's contribution to vowel identification can be reduced by mistuning its frequency so that it is not a multiple of the fundamental, indicating that small changes in the acoustic properties of sounds can impact their higher-level phonetic percept (Darwin & Gardner, 1986; Darwin & Sutherland, 1984).

The sequential organization of speech also requires grouping abstract linguistic units. In particular, phonemes must be grouped into words, which usually includes the additional step of segmenting the speech stream at multiple time points to yield distinct perceptual units (i.e., words). Formant transitions have been shown to assist with grouping by serving as the "glue" that helps bind adjacent phonemes (Cole & Scott, 1973; Stachurski, Summers, & Roberts, 2015). Familiarity with the spoken sequence also increases cohesion among phonemes. For example, phonemes that constitute words cohere more strongly than those that make up nonwords (Billig, Davis, Deeks, Monstrey, & Carlyon, 2013). The process of

✉ Marjorie Freggens
freggens.1@osu.edu

[1] Department of Psychology, Ohio State University, Columbus, OH 43220, USA

segmenting continuous speech into words is thought to be driven by a complex interplay of lower-level and higher-level language processes (Mattys, White, & Melhorn, 2005). Although acoustic-phonetic cues at word boundaries appear to take precedence in segmentation, contextual information (lexical, sentential) is used to aid segmentation when such cues are ambiguous or absent (Kim, Stephens, & Pitt, 2012; Mattys & Melhorn, 2007).

Auditory grouping processes are assumed to take precedence over higher-level linguistic influences (Bregman, 1990; Darwin, 2008; Shinn-Cunningham & Wang, 2008), but little is known about how language processes per se contribute to sequential grouping in speech. Remez, Rubin, Berns, Pardo, and Lang (1994) take the view that grouping processes for speech are dedicated to binding together the acoustic patterns of phonemes into a speech stream. Our goal was to inform thinking on these issues by studying whether knowledge of language influences how listeners organize the incoming speech signal over time, in addition to its role in aiding perception of the talker's message (e.g., linguistic units such as phonemes and words).

There is some evidence that the perceptual organization of speech is influenced by one's linguistic experience. For example, Best and Avery (1999) examined hemispheric differences in how monolingual English speakers perceived Zulu clicks. The typical left-hemisphere advantage found with familiar phonetic sequences was absent in English speakers when listening to click consonants, suggesting that experience with the phonetic inventory of a language not only determines the comprehension of the sounds, but also the perceptual processing that can be applied to the sound by the listener. Similarly, Yoshida et al. (2010) explored the biases of 5- to 6-month and 7- to 8-month English- and Japanese-learning infants in grouping rhythmic sequences of tones. Tones were presented as trochees (long-short) and iambs (short-long). The 5- to 6-month-old group showed no difference in biases between rhythmic classifications. However, the 7- to 8-month-old group showed a difference wherein English learning infants were more likely to perceive the tones as iambs. Such evidence illustrates how language experience can shape the way in which listeners organize sounds.

We explored linguistic influences in the sequential grouping of speech by dissociating auditory grouping processes from those attributable to knowledge of the language. This was accomplished by comparing how listeners grouped (segmented) two-word sequences when heard as nonspeech and then when heard as speech. Differences in performance across these two listening conditions should identify ways in which linguistic knowledge affects sequential grouping over and above those due to auditory processes alone.

In casual speech, fast articulation causes word boundaries to become blurred because of co-articulation between the final phoneme of one word and the initial phoneme of the next. When these two adjacent phonemes are the same, the talker may produce a slightly elongated, single phoneme that can be thought of as a fake geminate (Oh & Redford, 2012), so named because a phonetic distinction based on duration is not made in English. An example is shown in the spectrograms of Fig. 1. Note that there is only a single [s] in the acoustic record where two would be expected if the word pair *loose seam* were articulated carefully.

The segmentation of such word pairs would seem to be a formidable grouping problem based on auditory processes alone. The perceived word boundary does not correspond to an obvious location in the acoustic signal. Rather, it is a product of language processes imposing a boundary where there is none in the signal. Use of such word pairs was intended to provide an environment that would make linguistic influences on grouping distinguishable from those due to auditory processes. When auditory and linguistic processes favor competing organizations, what is the nature of their dynamics that determines the interpretation experienced by the listener? Exploration of this question can help delineate how and to what degree linguistic processes influence perceptual organization.

Experiment 1

Listeners responded to sinewave analogs of the word pairs presented in separate blocks of trials, first when informed they were sequences of nonspeech sounds and then again after being informed the stimuli were speech. The nonspeech and speech stimuli were acoustically quite similar but not identical (see Method). A cycling paradigm (Bregman & Campbell, 1971; Warren, 1968), in which a stimulus is repeated without pause for a short time, was used to measure perceptual (re)organization. Initially, the veridical stimulus is perceived, but after a few repetitions, the stimulus may be reorganized into another percept that consists of a foreground object that forms the focus of one's attention and a secondary, background object. Participants' task was to isolate the first foreground percept that was heard, which was accomplished by moving two "cursors" (one at the start of the sound file and one at the end), so that only the foreground was audible (Pitt & Shoaf, 2002). The cycling paradigm is thought to tap early sequential grouping processes, because reorganization of the percept is considered to be driven largely by the acoustic properties of the stimuli, especially at fast repetition rates (Diehl, Kluender, & Parker, 1985; Stachurski et al., 2015; Warren & Gregory, 1958).

When heard as nonspeech, reorganization is likely to occur at points in the signal where acoustic continuity is weak (Warren, Obusek, Farmer, & Warren, 1969). Inspection of the spectrogram at the top of Fig. 1 identifies two such locations, at the onset and offset of [s] frication, where there are

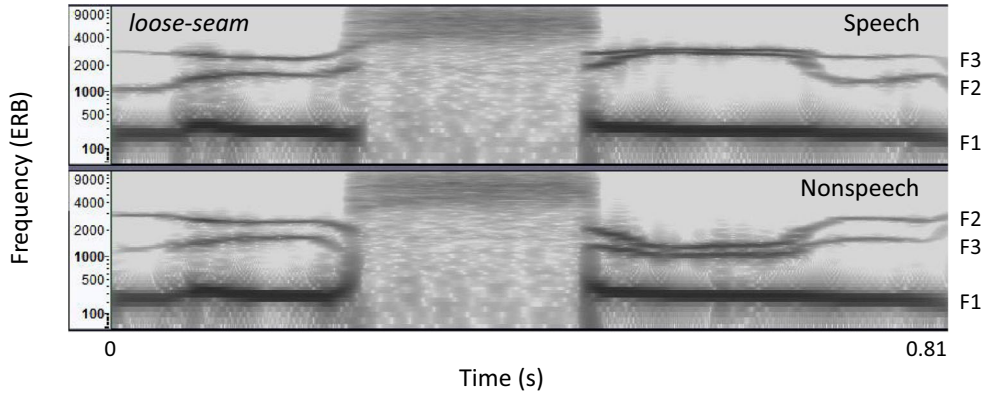

Fig. 1 A spectrogram of a sinewave stimulus, "loose-seam," in the speech condition (**top**) and in the nonspeech condition (**bottom**). The y-axis is ERB scaled to make the formants visible. See text for additional details

abrupt changes on multiple acoustic dimensions (periodicity, amplitude, frequency centroid, and bandwidth) when transitioning into and out of the [s]. It is at these points that the stimulus is likely to reorganize perceptually (Bregman, Colantonio, & Ahad, 1999). The [s] could break off and form its own grouping or cohere to one of the adjacent tonal portions. Cursor placement in this condition could include one of these locations depending on how [s] binds with the surrounding context.

How might knowledge that the stimulus is speech alter reorganization? Pitt and Shoaf (2002; see also Billig et al., 2013) presented monosyllables to listeners in a similar experimental setup and observed differences in cursor placement as a function of whether the sinewave analogs were heard as speech. In their speech condition, cursors tended to be placed close to phoneme boundaries, suggesting that perceptual organization was influenced by the phonetic content of the percept. Similar influences here could manifest themselves in a few ways. Linguistic influences could have a strong impact on grouping, in which case the two-word percept would be resistant to re-organization and listeners would hardly move the cursors, leaving them at the start and end of the sound file. If, instead, the two-word percept is a weak perceptual object, then re-organization should occur. Because the stimuli are two words that share an [s] at their boundary, there is no obvious *acoustic location* at which to place one of the cursors, but lexically there is reason to split the [s] in half, with the other cursor being placed at the beginning or end of the sound file depending on whether the first (e.g., *loose*) or the second (e.g., *seam*) word formed the foreground. This outcome would reflect weaker linguistic influences on grouping; the initial percept is not resisting reorganization, but is still being organized into a linguistic percept. If instead regrouping of speech is governed by the same auditory process as in the nonspeech condition, cursors would be placed in approximately the same location as in the nonspeech condition, at the start or end of the [s], where discontinuities in the signal are greatest.

Method

Stimuli

Target stimuli were sinewave analogs of 16 two-word sequences (e.g., *loose seam*, *plus sign*; mean length=710 ms; stimuli are listed in the Appendix). They were generated from natural tokens (recorded by a male speaker) using an LPC-based approach in Praat (Boersma & Weenink, 2018). The program first generated an initial set of frequency trajectories for the first three formants. To improve intelligibility and correct for potential errors, formant trajectories were then manually adjusted based on visual inspection of speech spectra prior to resynthesis for all items. In addition, to ensure clear /s/ perception, we replaced the sinewave [s] in each token, the acoustic characteristics of which make it a poor token of /s/,[1] with the [s] from the natural production of the word pair. Coarticulatory cues from the surrounding vowels that were present in [s] were preserved by including the natural [s], causing these hybrid stimuli to sound more intelligible than without the natural [s]. The abrupt change in signal characteristics from sinewaves to wide-band noise is not perceptually jarring. Example stimuli can be found on the authors' website.

Pilot testing in which we assessed the ability to bias perception of the stimuli as nonspeech (e.g., using instructions and filler trials of tone-noise sequences) failed. All listeners immediately reported hearing the stimuli as speech. We therefore had to modify the stimuli to avoid participants spontaneously hearing them as speech in the nonspeech condition. Nonspeech versions were created by inverting the second and third (sinewave) formants about their mean frequency (inverting either alone was insufficient). This manipulation

[1] Sinewave versions of fricatives like /s/ do not accurately depict the aperiodic noisiness of such sounds; the active frequency bands from the formants carry through the sinewave fricative, causing it to sound like a high-pitched vowel. Replacement of the sinewave /s/ with the natural [s] token for that item improved stimulus intelligibility.

has been used by others (Blesser, 1972; Humphries, Sabri, Lewis, & Liebenthal, 2014; Roberts, Summers, & Bailey, 2010; Scott, Blank, Rosen, & Wise, 2000; Vandermosten et al., 2010) to achieve the same purpose; it eliminates much of the phonetic quality of the stimuli while simultaneously keeping their spectro-temporal properties very similar to the unaltered versions. An additional eight word-pairs without medial [s] (e.g., *lean-in*) served as practice and filler items, and were processed in the same way as the target stimuli.

The intelligibility of the speech and nonspeech stimuli were evaluated in a pilot study in which participants (N=19) listened to each sinewave token and reported the percept. When listening to the nonspeech stimuli, participants reported a nonspeech sound or acoustic description (82%) more than a vocal description (speech or unspecified voices: 18%). When listening to the speech stimuli, the same participants reported speech 100% of the time, and their reports were the intended two-word percept 79% of the time.

Procedure

A custom Python script controlled stimulus presentation and response collection. Stimuli were presented over headphones. On each trial, the stimulus repeated without pause. Participants were instructed to first verbalize what they perceived the entire stimulus to be, and then to listen until a foreground (attended portion) was perceived. They were then to isolate (window) the start and end of the foreground using keys on the computer keyboard. Two buttons moved the start of the window earlier or later in time, and two buttons moved the end of the window earlier or later in time. These auditory cursors were initially positioned at the start and end of the sound file, respectively. Every button press moved the corresponding cursor 20 ms. The repeating stimulus was immediately updated, without pause, to exclude or include more of the sound file depending on the direction the cursor was moved. The shortest interval allowed between the start and end of the window was 100 ms. After participants isolated the foreground, they pressed the Enter key to stop the stimulus from repeating and verbalized the percept that they had isolated. Trials were self-paced in that participants had as many repetitions as needed to isolate the foreground. Participants initiated the next trial when ready.

The experiment was blocked by listening condition, with participants performing the windowing task first with the nonspeech stimuli and instructions (e.g., computer bleeps and noises) and then again with the speech stimuli and instructions. Each block contained 16 trials (12 test and four practice). Of the 12 test trials, eight trials were target trials and four were filler trials. No stimuli were repeated from the nonspeech block to the speech block, but stimulus lists were counterbalanced across groups of participants so that all stimuli occurred equally in the speech and nonspeech conditions

Participants

Fifty-two Ohio State University undergraduates participated in exchange for course credit. All were native English speakers with self-reported normal hearing. The data from 23 participants were excluded from analysis because of reporting speech (e.g., vowels, words, voices) on *any* trial in the nonspeech condition. We used a conservative exclusion criterion to ensure that data in the nonspeech condition did not reflect the operation of speech processes. Only the data from the remaining 29 participants were analyzed in this experiment. A subset of the data from the excluded participants was analyzed, and found not to be qualitatively distinct from the included participants. The final sample size was justified by performing a power analysis with an effect size of .4 at an alpha level less than .01. With 29 participants, we should be able to detect an effect with a power of .92 (G*Power: Faul, Erdfelder, Lang, & Buchner, 2007).

Results and discussion

In the nonspeech condition, participants reported mostly computerized tones or alarm sounds. In the speech condition, participants perceived the word pairs correctly on 78% of the trials; only responses on these trials were used in the analyses. Verbalization of the percept that was heard when cycling started took longer for the nonspeech stimuli (M=22 repetitions, ~15 s) than for the speech stimuli (M=11 repetitions, ~7 s). In addition, participants took more than twice as many repetitions to identify and window the foreground in nonspeech (M=71) than in speech trials (M=32). These results point to phonetic and lexical knowledge facilitating responses to the cycling speech.

Inspection of the cursor placement data showed that participants windowed the stimuli in one of two ways 87% of the time. In the first (40%), the foreground started at stimulus onset, and included some or all of [s], sometimes more. In the second (47%), the foreground started near [s] onset and ended at stimulus offset. We therefore separated responses as a function of which section (denoted fg1 and fg2) formed the attended foreground and analyzed differences in cursor placement as a function of listening condition in each of the two foregrounds. The [s] was excluded from the foreground in 9% of the responses, and an additional 4% of the responses included almost the entire stimulus among other responses. The small proportion of responses that resisted regrouping (i.e., when participants windowed the entire stimulus) indicates that linguistic influences were rarely sufficiently strong to prevent regrouping of the stimulus.

The upper two horizontal lines in Fig. 2 are graphs that depict cursor placement in the stimulus. Each short vertical tick represents a cursor position by a participant, and has been time-adjusted to the average stimulus duration; those in the

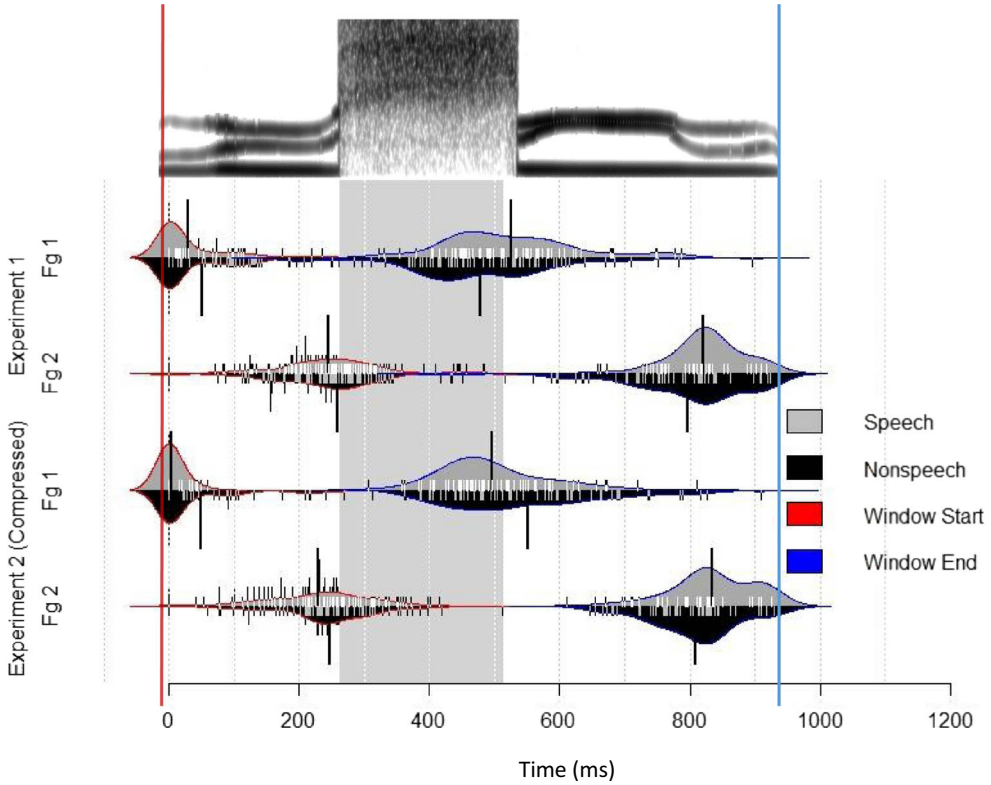

Fig. 2 The spectrogram of the sine-wave stimulus "loose seam" is at the top, and is time aligned along the x-axis with the cursor-placement graphs below. Short vertical tick marks denote placement of the start and end cursors in the windowing task, and are time-normalized across stimuli. Fg1 and fg2 refer to whether the foreground corresponded to the first or second part of the stimulus. The gray and black distributions depict the density of responses in the speech and nonspeech conditions, respectively. Distributions that are outlined in red show the start of the foreground window, and those in blue show the end of the foreground window. The long black vertical bars show the mean of each distribution. The vertical red line indicates the start of the stimulus, and the vertical blue line the end of the stimulus. The thick gray vertical bar denotes the position and duration of [s] averaged across stimuli. Its slight misalignment with the [s] above it is due to the example not matching exactly the onset and duration of the average [s]

speech condition are above the horizontal line, and those in the nonspeech condition are below it. Gray (speech) and black (nonspeech) density estimation functions are overlaid on the raw data to aid visual comparison of performance in the two conditions (Kampstra, 2008). The long black vertical ticks denote the means of the distributions. The thick gray vertical region in the middle of the graph denotes the average position of [s]. A spectrogram is positioned above the graphs to help visualize how the foreground placement data maps onto the stimuli.

Looking first at the fg1 responses, the distributions of cursor placements in the two listening conditions are similar in shape and location. The start distributions have peaks near acoustic onset and a right tail that extends about 100 ms into the token. The end distributions are much wider and straddle [s] offset, as indicated by the locations of the means. Inferential statistics were computed using the R statistical package (R core Team, 2013). A two-way repeated measures ANOVA on cursor placement with listening condition (nonspeech, speech) and cursor (start, end) as factors yielded a main effect of cursor, $F(1,66)=967.42$, $p=.001$ (no main effect of listening condition: $F(1,66)=1.22$, $p=.27$), showing that the end cursor was placed on average 465 ms after the start cursor. An interaction between listening condition and cursor was close to reaching significance, $F(1,66)=3.5$, $p=.065$, showing that the speech end boundary was placed 44 ms later than the nonspeech end boundary, and the speech start boundary was placed 11 ms earlier than the nonspeech start boundary. Thus, for the speech stimuli more of the acoustic material cohered, and was included in the foreground, whereas onset and offset phonemes were more likely to be segregated from the foreground when heard as nonspeech. For example, the plosive accompanying the onset of *plus* in *plus sign* is easy to integrate when heard as speech (cf. Best & Avery, 1999), but the same kind of burst presented as nonspeech is often perceived as separate from the periodic portion of the stimulus.

When the foreground was selected to be in the second half of the token (fg2), the outcome was similar except that the distributions shifted rightward, later in time. The start distribution for both listening conditions straddles [s] onset, and the end distribution, which is now more peaked, is centered 190 ms before file offset. A two-way ANOVA yielded only

a main effect of cursor, $F(1,84)=1790.32$, $p=.001$ (listening condition: $F(1,84)=.12$, $p=.71$), with the end cursor being on average 561 ms after the start cursor. The proclivity for the window to be wider in the speech than the nonspeech conditions was not present in fg2, $F(1,84)=.77$, $p=.38$.

The similarity of the windowed sections in the speech and nonspeech conditions in fg1 and fg2 suggests that perceptual grouping in the two conditions was quite similar, even though the percepts themselves were quite different. Indeed, in the speech condition, the foreground was reported as one of the two words (e.g., *loose* or *seam*), whereas in the nonspeech condition, the foreground was reported as a tone or whistle sound. Cursor placement across the conditions differed only in that a larger section was windowed when heard as speech in fg1, possibly to ensure complete inclusion of a boundary phoneme. Cursor placement was not always precise, as shown by some participants placing the cursor in the middle of the [s] and some beyond the end of the [s], but this behavior was similar in the two listening conditions. Auditory processes were likely the dominant force in sequential grouping in both listening conditions, as the similar distributions of cursor locations show. Linguistic knowledge affected only the interpretation of the foreground, hence the slight difference in mean cursor placement.

Linguistic influences in sequential organization might also manifest themselves in how [s] binds with the adjacent syllable. In the speech condition, [s] is a phonetic object, which could exhibit grouping biases that differ when heard as a noise. In particular, listeners tend to syllabify words such that consonants form the onset of a following syllable rather than the offset of the preceding one (Eddington, Treiman, & Elzinga, 2013; Treiman, 1992; see also Clements, 1992). Such a bias would manifest itself as a tendency to report [s] as the onset of the fg2 rather than the offset of fg1. No such bias should be found when [s] is heard as noise.

We found a strong tendency for [s] to be grouped with the rest of the foreground (91% of all responses), but where in the foreground was the [s] attached? The left half of Table 1 shows the proportion of [s]-included foreground responses in the speech and nonspeech conditions as a function of whether [s] grouped with fg1 or fg2. There is an affiliation bias for fg2 across both listening conditions, but it is more than three times greater in the speech condition. When heard as nonspeech, there was a slight (0.04) bias for [s] to form the end of the foreground. When heard as speech, the bias was 0.14. A chi-squared test for the interaction between listening condition and foreground was reliable, $\chi^2(4)=4$, $p=.04$. Comparisons within each listening condition showed the fg2 advantage to be reliable in the speech condition, $\chi^2(2)=13.8$, $p=.001$, but not in the nonspeech condition, $\chi^2(2)=.52$, $p=.46$. This [s] affiliation bias may reflect phonological influences in perceptual grouping. Word frequency, a likely alternative explanation, cannot account for this difference because the frequencies of the first and second words were not reliably different (p =.89).

Table 1 Proportion of [s]-included foreground responses in the speech and nonspeech conditions as a function of whether [s] grouped with fg1 or fg2

	Experiment 1		Experiment 2	
	Nonspeech	Speech	Nonspeech	Speech
fg1	.22 (73)	.19 (57)	.18 (64)	.20 (72)
fg2	.26 (82)	.33 (104)	.20 (70)	.28 (101)

Frequency counts are in parentheses

Proportions do not add up to one in each experiment because other types of responses were sometimes generated

The results of Experiment 1 provide mixed evidence of linguistic influences in the sequential organization of speech. On the one hand, the initial grouping of the input as two words containing two /s/ phonemes was overridden by auditory grouping processes, and led to the formation of a foreground percept that was acoustically similar across listening conditions. However, hearing the sole medial [s] phonetically, as opposed to hearing it as a noise, caused a marked change in grouping at a foreground boundary that is suggestive of linguistic influences.[2]

Experiment 2

The [s]-affiliation bias observed in Experiment 1 was discovered while exploring differences in verbal reports between the two listening conditions. In Experiment 2, we harnessed this result to test linguistic influences on perceptual grouping further. Diehl et al. (1985; see also Dorman, Cutting, & Raphael, 1975) found that repeating syllables (/s/-stop-vowel and stop-vowel pairs) streamed more as the repetition rate increased. They interpreted this result as suggesting that faster repetition rates reduce linguistic influences on phonetic perception (via grouping processes). Applying this same reasoning here, we increased the repetition rate of the word pairs with the expectation that the [s]-affiliation bias in the speech condition would shrink. Such an outcome would suggest that linguistic influences are modulated by the strength of auditory grouping processes. On the other hand, a similarly-sized [s]-affiliation bias to that found in Experiment 1 would highlight the persistence of linguistic influences in perceptual organization.

[2] To alleviate concerns that the fg2 affiliation bias we found was due to stimuli differences between nonspeech and speech stimuli (inverted f2 and f3 formants in nonspeech), we ran a pilot study where the nonspeech stimuli were time-reversed versions of the speech stimuli. This manipulation preserved the acoustic continuity of [s] with the surrounding formants while reducing their phonetic quality. If the fg2 bias we found in the speech stimuli was due to continuity issues in [s] onset, then time-reversing the speech should have produced a bias with fg1 instead. We did not find such a bias in the data.

Method

The experiment was identical to Experiment 1, except that the stimuli were compressed to 75% of their original duration (using a PSOLA algorithm: Hamon, Mouline, & Charpentier, 1989), a value identified in a pre-test to yield an 80% level of intelligibility for the word pairs. Fifty-two new individuals from the same pool as Experiment 1 participated. Following the exclusion criteria of Experiment 1, the data from 29 participants (those who did not report speech in the nonspeech condition) were included in the analyses.

Results and discussion

Participants' initial reports of the stimuli were veridical in 77% of the speech trials, similar to Experiment 1. No participants in the nonspeech condition spontaneously heard the stimuli as words, although 13% of responses included reports of computerized voices. These trials were dropped from the analysis as a precaution against including any responses that might be speech-based, although their inclusion does not change the results in a meaningful way. As in Experiment 1, fg1 (42%) and fg2 (51%) responses dominated, constituting 93% of all responses. The [s] was excluded from the foreground in 15% of the responses, and an additional 5% of the responses included almost the entire stimulus; no interesting patterns was found in these minority responses or in their comparison with those from Experiment 1.

As might be expected with a faster presentation rate, initial reports of the percept and windowing the foreground took longer than in Experiment 1, but the pattern of results was the same. Verbalization of the initial response required more repetitions for nonspeech trials (M=24) than speech (M=13), and participants took twice as many repetitions to window the foreground in nonspeech trials (M=71) than in speech trials (M=36). Here again listeners leveraged their linguistic knowledge to organize the perceptual stream quickly.

The cursor data were scored and analyzed in the same way as in Experiment 1, and are shown in the bottom two graphs of Fig. 2. The cursor values were adjusted to the time scale of the top two graphs to facilitate comparison with the data from Experiment 1. Initial comparison of the cursor-placement distributions across experiments shows a similar outcome to that with the time-compressed stimuli. Cursor placement distributions are similar in the speech and nonspeech conditions. In the fg1 data, statistical comparisons yielded a main effect of cursor, $F(1,68)=866.19$, $p=.001$, such that end cursor placement was 506 ms after start cursor placement. In addition, a slight effect of listening condition was found, $F(1,68)=5.66$, $p=.02$, such that speech cursors were placed 41 ms earlier than nonspeech cursors. This is a reversal of the effect found in Experiment 1. We as yet do not have an explanation for it, other than to note that mean location is somewhat unstable given the wide variability in cursor placement at [s] offset. Unlike in Experiment 1, there was no interaction between these variables ($F(1,68)=.11$, $p=.74$). A similar pattern was observed in the fg2 data: main effects of cursor, $F(1,82)=2113.75$, $p=.001$ and listening condition, $F(1, 82)=4.57$, $p=.03$ were obtained (interaction: $F(1,82)=2.76$, $p=.1$), with the end cursor placed 581 ms after the start cursor, and speech cursors placed 28 ms before nonspeech cursors. The presence of an interaction of listening condition with cursor placement in Experiment 1, and the subsequent muting of that effect in Experiment 2, is what is expected if increasing the repetition rate reduces linguistic biases in sequential integration.

The right side of Table 1 shows the frequency with which [s] grouped with fg1 and fg2 in the two listening conditions. The faster presentation rate caused biases in [s] grouping to weaken. In the nonspeech condition, [s] now groups with fg2 only slightly more than fg1, a bias of (.02). In contrast, the asymmetry found in speech condition of Experiment 1 persisted in Experiment 2, but it is almost half the magnitude (.14 vs. .08). A chi-squared test showed that the interaction of listening condition and foreground in Experiment 2 was not reliable, $\chi^2(4)=.92$, p=.34. When analyzed alone, the foreground difference in the speech condition was reliable, $\chi^2(2)=4.86$, $p=.03$, but the difference in the nonspeech condition was not, $\chi^2(2)=.27$, $p=.6$. This outcome suggests that the faster stimulus presentation rate reduced linguistic influences on grouping, but did not remove them entirely. The drop in the magnitude of the [s]-bias effect in the speech condition across experiments was reliable, $\chi^2(2)=4.26$, $p=.04$.

The results of Experiment 2 replicate and build on those of Experiment 1. A faster stimulus repetition rate further reduced the differences in cursor placement across listening conditions, and reliably diminished by almost half the tendency of [s] to form the onset of the foreground in the speech condition. The faster presentation rate had the intended effect of lessening linguistic influences on sequential grouping, demonstrating both the primacy of auditory processes in sequential organization and the influence of linguistic knowledge in integrating the pieces that constitute a speech object.

General discussion

Do linguistic processes aid in the sequential organization of speech over and above their role in facilitating message understanding? We addressed this question by comparing how listeners grouped repeating two-word sequences when heard as (slightly altered) nonspeech and then again as speech. Of interest was how grouping of [s] changed when heard as a phonetic segment versus a hiss. The results across two experiments provide evidence of linguistic influences, and also identify constraints on such influences imposed by auditory grouping processes.

Veridical perception of the two-word sequences, in particular that they contained two consecutive /s/ phonemes, shows that linguistic processes (e.g., lexical knowledge) can yield unique perceptual groupings of the speech signal, ones for which the grouping is not obvious given the acoustic signal. However, the fragility of this organization was exposed in the cycling paradigm, where repetition of the two-word sequences caused the initial grouping to reorganize into a foreground object that was acoustically quite similar to that formed with the nonspeech stimuli. Listening condition qualitatively changed interpretation of the foreground, but only minimally impacted the windowed region that defined the foreground, stretching it slightly in the speech condition in Experiment 1. This outcome suggests that even under circumstances benefitting greatly from linguistic knowledge (perceptual ambiguity, impoverished signal), auditory grouping processes constrain, and in some cases dominate, perceptual organization.

Drawing conclusions about perceptual organization from the cursor data across listening conditions involves making claims from a null result. Although replication of the null effect across two experiments strengthens the claim, we carried out an additional analysis to test it directly. The data were combined across experiments to increase the power of the analysis, and we quantified the evidence in favor of no difference between listening conditions (null hypothesis) by performing a repeated measures Bayesian ANOVA using the JASP software package (JASP Team, 2018; version 0.9). Default, diffuse priors were used. The effect of listening condition yielded a Bayes Factor of 5.39 (i.e., the null is 5.39 times more likely than the alternative), which is moderate evidence (Wagenmakers et al., 2018) for no difference in cursor placement between the speech and nonspeech conditions. Similar results were obtained when the same analysis was performed separately on the data in each experiment. This outcome further strengthens the claim that auditory processes constrain the influence of linguistic knowledge.

The interplay between linguistic and auditory grouping processes was observed by varying the cycling rate of the stimuli. In the speech condition of Experiment 1, [s] was reported as the onset of fg2 almost twice as often than as the offset of fg1. The reduction by half in the magnitude of this bias at a faster presentation rate (Experiment 2) suggests linguistic biases (e.g., phonological preferences to begin an utterance with a phoneme that is low in sonority; Clements, 1992) were dampened when auditory grouping forces were stronger.

The current results fit with the received view that linguistic influences occur downstream of general auditory grouping processes in speech perception (Bregman, 1990; Darwin, 2008; Sussman, 2017; Shinn-Cunningham & Wang, 2008). However, it is not that auditory processes always override higher-level linguistic influences to fully determine how speech groups sequentially. Rather, the results suggest that they constrain linguistic influences, and the cycling paradigm is one means of exposing these constraints. In particular, they suggest that sequential groupings that are defined by solely linguistic processes (perception of double /s/), with little corresponding support in the acoustic signal, are fragile, occurring only under specific conditions. In contrast, those that align with groupings for which there is acoustic evidence are more robust. This was the case in the [s] affiliation analysis, where [s] onset and offset corresponded with acoustic junctures. Indeed, that the bias to perceive [s] as an onset did not disappear completely at the fast repetition rate (Experiment 2) suggests that under some conditions, linguistic influences can be resistant to auditory regrouping forces, although it would be reasonable to expect they would disappear at an even faster rate. The persistence of linguistic influences could be indicative of well-learned stimuli. Familiarity has a strong influence on the grouping of melodies (Devergie, Grimault, Tillmann, & Berthommier, 2010; see also Sussman, 2017), suggesting that extensive exposure can have a significant influence on grouping (Moore & Gockel, 2012).

The above interpretation leads to questions about the architectural relationship between processes involved in perceptual grouping and those responsible for language understanding. Specifically, how do the two interface? Although the current data are suggestive of a feed-forward relationship, with auditory grouping processes preceding higher-level linguistic ones, consideration of the many time scales at which both processes operate suggests a more complex structural organization (Mill, Bőhm, Bendixen, Winkler, & Denham, 2013). Sequential grouping processes operate over a range of time scales, organizing the auditory scene across fractions of a second (isolating the direction of a voice) to many seconds and beyond (e.g., tracking a melody). Spoken language consists of units that span multiple time scales, too, from features and phonemes to words, sentences, and turns in conversation. Instead of grouping processes feeding into linguistic processes at a single juncture, the two processes may inform each other across these time scales, possibly defined by the size of the linguistic unit itself.

Examples of such interactions across time scales are hinted at in the literature. The identification of individual phonemes has been shown to be influenced by auditory grouping, as the stimulus [ska] is perceived as [s] and [ga] when repetition causes the [s] to stream (Diehl et al., 1985). The phenomenon of phonological fusion, in which phonemes combine across ears to create blended words (*back (left ear)* + *lack (right ear)* = *black*; Cutting, 1973a, Sexton & Geffon, 1981) is an example of how language constrains the sequencing of phonemes.

As mentioned in the *Introduction*, knowledge of words increases the cohesion of their phonemes (Billig et al., 2013; Pitt & Shoaf, 2002). Knowledge of articulation may also guide the grouping of sinewaves when they are in competition (Roberts et al., 2015; Summers et al., 2016). Word segmentation requires the assignment of acoustic information to the correct word at word junctures, which in continuous speech can be ambiguous (Ernestus, Baayen, & Schreuder, 2002; Gow, 2003), as the stimuli in the current experiment illustrate. Although grouping and linguistic processes operate independently and serve different functions in perception, in the context of spoken language their interaction facilitates organization and encoding of the speech signal.

Caveats on interpretation of the current data are of course in order. The assumption that the nonspeech results reflect early auditory grouping while the speech results reflect higher-level influences on perceptual organization might be too strong. It could be that perceptual organization is taking place at a single level of analysis, and what differs is how that information is organized based on the listener's attentional set, listening to speech versus nonspeech (Bregman, 1990). The task of reporting the foreground percept, the end result of perceptual organization, might have yielded data that reflect only later stages of organization, and not grouping at early auditory stages. The current study assumes auditory grouping to be involved in word segmentation. One might take issue with this stance and consider segmentation solely the domain of higher-level language processes. Parsing the acoustic signal at word boundaries is a complex problem that requires integrating multiple sources of information. It is difficult to understand how segmentation would not be influenced by auditory grouping when the cohesion of speech itself has been shown to be sensitive to such influences (Ciocca, 2008; Nygaard, 1993). There is a great deal we do not know about auditory perceptual organization, including how it interfaces with other aspects of perception and cognition.

In sum, the centrality of language in human functioning has led researchers to devote considerable attention to elucidating its underlying mechanisms (Diehl, 2008; Diehl, Lotto, & Holt, 2004). The current study expands what it means for knowledge of the language to aid speech perception. In addition to facilitating phonetic encoding, it is recruited to ensure the speech signal itself coheres over time. Although conceptually phonetic perception and perceptual grouping can be thought of as different problems, they operate to achieve the common goal of aiding the listener in understanding the talker's message.

Acknowledgements We thank Sydney Sivier, Karin Luk, and Alyssa Sabo for their contributions in collecting and scoring data.

Appendix 1: List of Stimuli

air-raid
boss-sent
brass-sword
class-song
close-sound
cross-site
dress-suit
face-sink*
gas-source
glass-sip
grass-seed
less-soap
line-up
lean-in
loose-seam*
main-way
more-rain
nice-side*
plus-sign
press-soft*
real-life
wear-out
wrong-lane
yes-such

*The items marked with an asterisk contain embedded words when the [s] groups only with the other syllable (e.g., loo –seam). These embedded words tended to be low in frequency, and removing them from the analysis did not change the results in a meaningful way.

References

Best, C. T., & Avery, R. A. (1999). Left-Hemisphere Advantage for Click Consonants is Determined by Linguistic Significance and Experience. *Psychological Science, 10*(1), 65–70. https://doi.org/10.1111/1467-9280.00108.

Billig, A. J., Davis, M. H., Deeks, J. M., Monstrey, J., & Carlyon, R. P. (2013). Lexical influences on auditory streaming. *Current Biology, 23(16)*, 1585-9.

Blesser, B. (1972). Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *Journal of Speech and Hearing Research*, 15(1), 5-41. doi: https://doi.org/10.1044/jshr.1501.05.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.40. http://www.praat.org/

Bregman, A. S. *Auditory Scene Analysis.* Cambridge, MA: MIT Press, 1990.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology, 89, 244–249.*

Bregman, A.S., Colantonio, C., and Ahad, P.A. (1999). Is a common grouping mechanism involved in the phenomena of illusory continuity and stream segregation? *Perception & Psychophysics 61,* 195-205.

Ciocca, V. (2008). The auditory organization of complex sounds. *Frontiers in Bioscience-Landmark*, *13*, 148–169. https://doi.org/10.2741/2666

Clements, G.N. (1992). The Sonority Cycle and Syllable Organization, *Phonologica 1988*, eds. W.U. Dressler, H.C. Luschützky, O. Pfeiffer, & J. Rennison, Cambridge University Press, Cambridge, 63-76.

Cole, R. A., & Scott, B. (1973). Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology, 27, 441–449.*

Cutting, J. E. (1973a). *Perception of speech and nonspeech, with and without transitions*. Haskins Laboratories Status Report on Speech Research, SR-33, Haskins Laboratories, (pp. 37–46).

Darwin, C. J. (2008). Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 363*(1493), 1011–1021. https://doi.org/10.1098/rstb.2007.2156

Darwin, C. J., & Gardner, R. B. (1986). Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. *Journal of the Acoustical Society of America, 79,* 838–845.

Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 193–208.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research, 229*(1–2), 132–147. https://doi.org/10.1016/j.heares.2007.01.014

Devergie, A., Grimault, N., Tillmann, B., & Berthommier, F. (2010). Effect of rhythmic attention on the segregation of interleaved melodies. *The Journal of the Acoustical Society of America*, *128*(1), EL1–EL7. https://doi.org/10.1121/1.3436498

Diehl, R. L. (2008). Acoustic and auditory phonetics: The adaptive design of speech sound systems. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *363*(1493), 965–978. https://doi.org/10.1098/rstb.2007.2153

Diehl, R. L., Kluender, K. R., & Parker, E. M. (1985). Are selective adaptation and contrast effects really distinct? *Journal of Experimental Psychology: Human Perception and Performance, 11* (2), 209-220.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*(1), 149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028

Dorman, M. F., Cutting, J. E., & Raphael, L. J. (1975). Perception of temporal order in vowel sequences with and without formant transitions, *Journal of Experimental Psychology: Human Perception and Performance, 104* (2), 121-129.

Eddington, D., Treiman, R., & Elzinga, D. (2013). Syllabification of American English: Evidence from a Large-scale Experiment. Part I. *Journal of Quantitative Linguistics, 20*(1), 45–67. https://doi.org/10.1080/09296174.2012.754601

Ernestus, M., Baayen, H., & Schreuder, R. (2002). The recognition of reduced word forms. Brain and Language, 81, 162–173.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191. https://doi.org/10.3758/BF03193146

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. Perception & Psychophysics, 65, 575–590.

Hamon, C., Mouline, E., & Charpentier, F. (1989). A diphone synthesis system based on time-domain prosodic modifications of speech. In *International Conference on Acoustics, Speech, and Signal Processing*, (p. 238–241 vol.1). https://doi.org/10.1109/ICASSP.1989.266409

Humphries, C., Sabri, M., Lewis, K., & Liebenthal, E. (2014). Hierarchical organization of speech perception in human auditory cortex. *Frontiers in Neuroscience*, *8*, 1-12. https://doi.org/10.3389/fnins.2014.00406

JASP Team. (2018). JASP (Version 0.9)[Computer software]. Retrieved from https://jasp-stats.org/

Kampstra, P. (2008). Beanplot: A boxplot alternative for visual comparison of distributions. Journal of Statistical Software, 28, 1–9. https://doi.org/10.18637/jss.v028.c01

Kim, D., Stephens, J. D. W., & Pitt, M. A. (2012). How does context play a part in splitting words apart? Production and perception of word boundaries in casual speech. Journal of Memory and Language, 66(4), 509–529. https://doi.org/10.1016/j.jml.2011.12.007

Mattys, S. L., & Melhorn, J. F. (2007). Sentential, lexical, and acoustic effects on the perception of word boundaries. The Journal of the Acoustical Society of America, 122(1), 554–567.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. Journal of Experimental Psychology: General, 134(4), 477–500. https://doi.org/10.1037/0096-3445.134.4.477

Mill, R. W., Bőhm, T. M., Bendixen, A., Winkler, I., & Denham, S. L. (2013). Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLoS Computational Biology*, *9*(3), e1002925. https://doi.org/10.1371/journal.pcbi.1002925

Moore, B. C. J., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *367*(1591), 919–931. https://doi.org/10.1098/rstb.2011.0355

Nygaard, L. C. (1993). Phonetic coherence in duplex perception. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(2), 268–286.

Oh, G. E., & Redford, M. A. (2012). The production and phonetic representation of fake geminates in English. *Journal of Phonetics*, *40*(1), 82–91. https://doi.org/10.1016/j.wocn.2011.08.003

Pitt, M.A., & Shoaf, L. (2002). Linking verbal transformations to their causes. Journal of Experimental Psychology: Human Perception and Performance, 28, 150-162.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. Psychological Review, 101, 129–156.

Roberts, B., Summers, R. J., and Bailey, P. J. (2010). The perceptual organization of sine-wave speech under competitive conditions. *Journal of the Acoustical Society of America, 128*, 804- 817.

Roberts, B., Summers, R. J., & Bailey, P. J. (2015). Acoustic source characteristics, across-formant integration, and speech intelligibility under competitive conditions. *Journal of Experimental Psychology: Human Perception & Performance, 41*, 680–691.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. Brain, 123(12), 2400–2406. https://doi.org/10.1093/brain/123.12.2400

Sexton, M. A. & Geffon, G. (1981). Phonological fusion in dichotic monitoring. *Journal of Experimental Psychology: Human Perception and Performance, 7* (2), 422-429.

Shinn-Cunningham, B. G., & Wang, D. (2008). Influences of auditory object formation on phonemic restoration. The Journal of the Acoustical Society of America, 123(1), 295–301. https://doi.org/10.1121/1.2804701

Stachurski, M., Summers, R. J., & Roberts, B. (2015). The verbal transformation effect and the perceptual organization of speech: Influence of formant transitions and F0-contour continuity. Hearing Research, 323, 22–31. https://doi.org/10.1016/j.heares.2015.01.007

Summers, R. J., Bailey, P. J., Roberts, B. (2016). Across-formant integration and speech intelligibility: Effects of acoustic source properties

in the presence and absence of a contralateral interferer. *The Journal of the Acoustical Society of America 140*(2):1227–1238.

Sussman, E. S. (2017). Auditory scene analysis: An attention perspective. *Journal of Speech, Language, and Hearing Research*, *60*(10), 2989–3000. https://doi.org/10.1044/2017_JSLHR-H-17-0041

Treiman, R. (1992). Experimental studies of English syllabification. In W. U. Dressler, H. C. Lushützky, O. E. Pfeiffer, & J. R. Rennison (Eds.), Phonologica 1988 (pp. 273–281). Cambridge, England: Cambridge University Press.

Vandermosten, M., Boets, B., Luts, H., Poelmans, H., Golestani, N., Wouters, J., & Ghesquière, P. (2010). Adults with dyslexia are impaired in categorizing speech and nonspeech sounds on the basis of temporal cues. *Proceedings of the National Academy of Sciences*, *107*(23), 10389-10394. https://doi.org/10.1073/pnas.0912858107

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Warren, R. M. (1968). Verbal transformation effect and auditory perceptual mechanisms. *Psychological Bulletin, 70*, 261–270.

Warren, R. M. (2008). Auditory perception: an analysis and synthesis. Cambridge, UK: Cambridge University Press.

Warren, R. M., & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *American Journal of Psychology, 71*, 612–613.

Warren, R.M., Obusek, C.J., Farmer, R.M., & Warren, R.P. (1969). Auditory sequence: Confusion of patterns other than speech and music, *Science 164*, 586-587.

Yoshida, K. A., Iversen, J. R., Patel, A. D., Mazuka, R., Nito, H., Gervaine, J., & Werker, J. F. (2010). The development of perceptual grouping biases in infancy: A Japanese-English cross-linguistic study. *Cognition, 115* (2), 356–361.