5-2022

Adjusting for speaking rate when perceiving speech in background noise.

Dawson C Stephens
University of Louisville

Adjusting for Speaking Rate when Perceiving Speech in Background Noise

Dawson C. Stephens

Honors Thesis

Honors Program

University of Louisville

February 23, 2022

Abstract

Speech perception is a very relevant concept occurring every day. Acoustic context effects such as temporal contrast effects (TCEs) influence perception significantly. For instance, when a faster context sentence is spoken, the participant should perceive the following target word as slower and more like /t/ in “tier”; when a slower context sentence is spoken, the participant should perceive the following target sound as faster and more like /d/ in “deer”. Recent work by Bosker et al. (2020) concluded that selective attention (directing attention to a specific stimulus while ignoring surrounding stimuli) had no effect on TCEs, suggesting they were automatic and low-level. However, their paradigm was not an ideal test; the voices heard contained different talkers with one presented to each ear, making them easy to perceptually separate. Here, the paradigm was designed to eliminate talker variability (acoustic variability among talkers) by using the same male talker speaking one sentence to both ears, two sentences simultaneously to both ears (diotically) or one to each ear (dichotically). Two experiments tested these effects of presentation mode on TCEs. In each experiment, TCE magnitudes were similar across presentation modes. These results are consistent with Bosker et al.’s (2020) claims of TCEs being automatic and low-level. Potential neural mechanisms contributing to TCEs are discussed.

Adjusting for Speaking Rate when Perceiving Speech in Background Noise

Everything we see, hear and do in life is based on our perception of the world around us. There are numerous different influences on our perception at any given time. According to Pardo et al. (2021), the ability to recognize spoken words is influenced by the talker, the listener, speech signal and context. In the absence of background noise, speech becomes easier for people to identify individual sounds; however, speech that is accompanied by background or environmental noise, including other speech, can be very difficult and at times, utterly impossible to distinguish (Pardo et al., 2021). One big contributor to this difficulty is the signal-to-noise ratio: as the noise becomes more intense relative to the speech, it becomes harder to understand (e.g., Miller & Nicely, 1955). Another contributor to this difficulty is how well the listener can separate the signal from the noise. In Brungart (2001), speech was best understood when the voices were easy to separate (i.e., one man and one woman talking simultaneously), more challenging when voices were harder to separate (i.e., two different men speaking), and most challenging when the two voices speaking belonged to the same person (and thus very difficult to distinguish the target message from the background noise).

Perception of speech sounds is based on intrinsic cues (i.e., acoustic properties of the sound itself) as well as extrinsic cues (i.e., acoustic properties of surrounding sounds, or the acoustic context; Ainsworth, 1975). Instances of extrinsic cues affecting speech perception are known as context effects. When perceiving speech, the sounds before or after the target sound form the context. There are two main types of acoustic context effects: Spectral Contrast Effects (SCEs; induced by variations in frequency) and Temporal Contrast Effects (TCEs; induced by speaking rate; Stilp, 2020). For the purpose of this experiment, TCEs were analyzed to test the effects of speaking rates and how speech perception was affected. For example, perception of a

consonant as /d/ (as in ‘deer’) or /t/ (as in ‘tier’) depends on its voice onset time (VOT), or how long it takes for the vocal cords to start vibrating when the sound is produced (/d/ has a much shorter voice onset time than /t/). Perception of the VOT of a sound is affected by the speaking rate of sounds (like a context sentence) spoken before it. When the preceding context sentence is spoken at a fast rate, the target sound is perceived as slower and more like “tier”; when the preceding context sentence is spoken at a slow rate, the target sound is perceived as faster and more like “deer” (e.g. Summerfield, 1981).

Context effects are commonly measured in quiet, but everyday perception is seldom in quiet. Imagine attending a sporting event, such as a basketball game. Not only do individuals perceive speech from other fans in the crowd, but also the sound of a bouncing basketball, the chanting from the cheerleaders, the referees’ whistles or even the sound of the buzzer. All of these sounds are being processed simultaneously, each affecting how an individual perceives speech. Investigating context effects in background noise is necessary in order to better understand how they contribute to everyday perception.

TCEs were studied amidst background noise by Bosker et al. (2020) by presenting different talkers simultaneously and varying where participants were instructed to direct their attention. Experiments 1 and 2 presented only one talker speaking a single context sentence at different rates (either slow or fast). Experiments 3-5 presented two different talkers per trial, and the participants were asked to focus their attention on a single talker. Experiment 6 was unique in that the participants heard two talkers simultaneously and were asked to focus on both talkers. Therefore, when only one talker was heard in Experiments 1-2, there was no competition for the participants’ attention; in Experiments 3-6, listeners heard two different talkers and were instructed to either attend selectively to one talker (Experiments 3-5) or divide attention across

both talkers (Experiment 6). Each experiment analyzed the Dutch morphological prefix /ge-/ forming the past participle of a present tense verb such as “gaan” and “gegaan”. Thus, fast contexts were predicted to make the target sound longer so the /ge-/ syllable was heard (making it past tense); slow context sentences were predicted to make the target sound faster so the /ge-/ syllable was not heard (making it present tense). All results turned out the same irrespective of attention instructions: when both talkers spoke at the same rate (either slow or fast), TCEs occurred; when talkers spoke at different rates (one speaking slowly and the other quickly), TCEs were extinguished. According to Bosker et al. (2020), selective attention did not change the effect of TCEs on target sound perception, suggesting they are automatic and related to relatively low-level processing.

Bosker et al. (2020) presented two talkers at the same time, but the talkers were two different women. Brungart’s (2001) study revealed that female voices can be perceptually separated to some degree improving overall speech perception. Also, the talkers were presented dichotically, one to each ear. Based on previous perception experiments, listeners excel at separating sounds when presented from different locations, known as “spatial release from masking” (Litovsky, 2012). These decisions allowed for participants in Bosker et al. (2020) to discriminate between simultaneous voices with little difficulty, so it was not a very strict test of their research question. A stricter paradigm would present the voices in the same location to remove the spatial release from masking effect and limit the speakers to being the same individual as this is the most difficult condition (Brungart, 2001).

These results constitute the remaining question: If matching speech rates influenced speech perception performance when perceiving two context sentences spoken by different talkers, then how would both context sentences being spoken by the same talker affect speech

perception? In this case, participants were expected to have difficulty separating the voices (as in Brungart, 2001), which may instead be perceived as a faster speaking rate (higher number of syllables per second). This may alter TCEs in a materially different way than was observed in Bosker et al. (2020). Two experiments were designed, each consisting of different pairs of context sentences (each pair matched in terms of syllable count and duration) to determine how individual versus simultaneous talkers affected TCEs in speech perception. Two experiments were designed because Experiment 1 produced a null result (as detailed below). To determine whether this finding was truly a null finding or a byproduct of the stimuli that were tested, new context sentences were selected and tested in an otherwise identical Experiment 2.

Methods

Participants

Twenty undergraduate students at the University of Louisville participated in Experiment 1, and twenty-two participated in Experiment 2. No listeners participated in both experiments. All students self-reported as native English speakers with normal hearing. These students participated in exchange for course credit in the Department of Psychological and Brain Sciences.

Stimuli

1. **Context Sentences**

In Experiment 1: Sentence 1 was a recording of an adult male saying, “Upgrade your status to reflect your wealth.” Sentence 2 was a recording of the same male saying, “What did you mean by that rattlesnake gag?”. The duration of each sentence was the same at 2098 ms with the same number of syllables (10). The rates of these sentences were edited using Praat

software (Boersma & Weenink, 2021) from 100% to 50% (duration divided by 2, altering their duration to 1049 ms) and from 100% to 150% (duration multiplied by 1.5, altering their duration to 3147 ms; Figure 1). Changing speaking rates by these amounts have successfully produced TCEs in previous experiments (Sharpe, 2021).

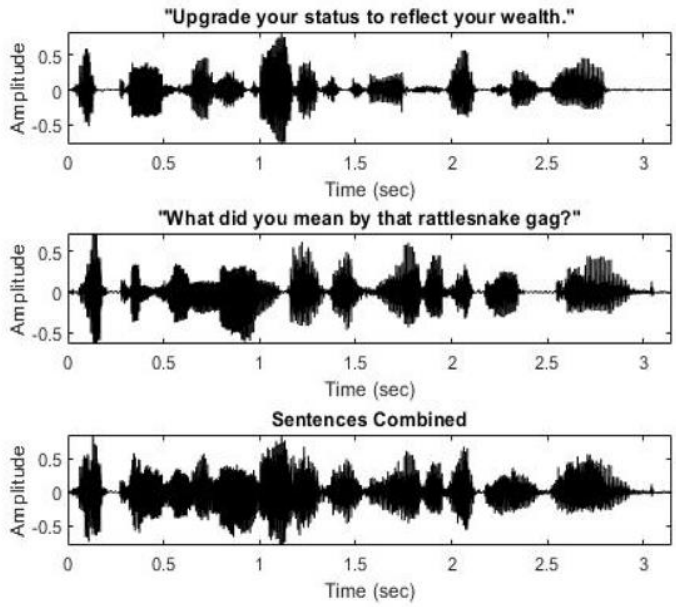

Figure 1: Experiment 1 context sentences displayed (slow speaking rates). Top: Sentence 1 waveforms shown in time (sec) and amplitude envelope. Middle: Sentence 2 waveforms shown in time (sec) and amplitude envelope. Bottom: Sentence 1 and Sentence 2 waveforms combined shown in time (sec) and amplitude envelope.

In Experiment 2: Sentence 1 was a recording of the same adult male mentioned previously saying, “The misprint provoked an immediate disclaimer.” Sentence 2 was a recording of the same male saying, “Even I occasionally get the Monday blues.” The duration of each sentence was the same at 2206 ms with same number of syllables (13). The rates of these sentences were edited using Praat software from 100% to 50% (duration divided by 2, altering

their duration to 1103 ms) and edited from 100% to 150% (duration multiplied by 1.5, altering their duration to 3308 ms; Figure 2).

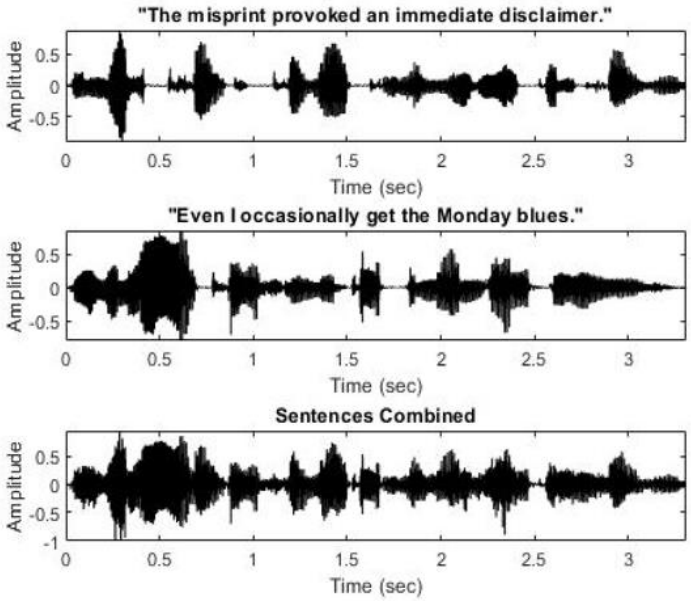

Figure 2: Experiment 2 context sentences displayed (slow speaking rates). Top: Sentence 1 waveforms shown in time (sec) and amplitude envelope. Middle: Sentence 2 waveforms shown in time (sec) and amplitude envelope. Bottom: Sentence 1 and Sentence 2 waveforms combined shown in time (sec) and amplitude envelope.

Targets

The target sounds were presented in a ten-step series (perceptually varied from "deer" to eventually sound more like "tier") varying from "deer" to "tier", based on recordings from the same adult male talker who spoke the context sentences. These stimuli were generated using Praat software by altering voice onset time (VOT), the duration of the consonant that is unvoiced before the vowel begins (Winn, 2020). The ten-step series consisted of the voiceless interval at the beginning of the target sound "deer" becoming longer until the target sound "tier" was

produced. Previous experiments have demonstrated that perception of these stimuli is sensitive to TCEs (Sharpe, 2021).

Procedure

Participants completed the experiments in the Auditory Perception and Processing Lab in the Department of Psychological and Brain Sciences. First, participants read and signed a consent form. Each participant wore headphones in a sound-attenuating booth. Before the main experiment, participants completed a practice block: they did 20 trials where the context was a neutral rate sentence, and the target was either the "deer" or "tier" endpoint of the 10-step series. Participants labeled the target word on each trial and received feedback. They were required to achieve at least 80% correct in the practice block before proceeding in the experiment, and all did. The main experiment consisted of four blocks containing 160 trials. Each block had 80 fast-context trials and 80 slow-context trials. Each of the 10-step series "deer"-"tier" targets were tested eight times following a slow sentence and eight times following a fast sentence. A trial consisted of one context sentence (slow or fast), or two context sentences (both slow or both fast) presented diotically (sentences added together in both ears) or dichotically (one sentence presented to one ear while a different sentence is presented to the other ear simultaneously) followed by a target sound ("deer" or "tier"). No feedback was provided.

Each experiment followed the same design. The blocks were tested in counterbalanced order, meaning each block appeared equally often in each position. One block consisted of fast and slow versions of Sentence 1, while another block consisted of fast and slow versions of Sentence 2. These two blocks represented the control condition for analyzing TCEs where there were no competing talkers (similar to Bosker et al., 2020). Another block was diotically organized (presenting both context sentences either fast or slow), while another block was

dichotically organized (presenting both context sentences fast or slow). All sounds were presented at a comfortable listening level of approximately 70 dB SPL (sound pressure level).

Results

The mean percentage of "tier" responses were calculated across the ten-step series. "Tier" responses to the target sound were predicted to be higher following fast sentences, so TCEs were calculated as percent "tier" responses following fast sentences minus percent "tier" responses following slow sentences (Sharpe, 2021).

TCEs were analyzed using one-way repeated measures ANOVA with condition (Sentence 1, Sentence 2, Diotic, Dichotic) as the independent variable (IV) and the percent shifts in "tier" responses as the dependent variable (DV).

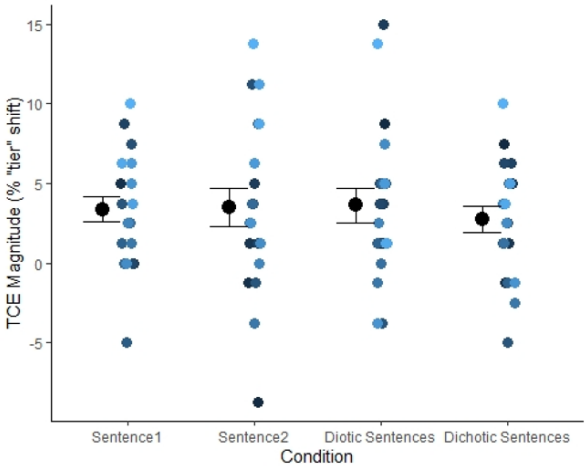

Figure 3: Experiment 1 percent "tier" shifts (TCEs) as a function of condition. Blue dots represent data from individual participants, and black circles depict condition means ± one standard error.

The mean TCEs and standard errors were calculated for each condition of Experiment 1: (Figure 3): Sentence 1 (*M* = 3.37%, *SE* = 0.8), Sentence 2 (*M* = 3.50%, *SE* = 1.2), Diotic (*M* =

3.63%, *SE* = 1.1), Dichotic (*M* = 2.75%, *SE* = 0.8). The one-way repeated measures ANOVA produced results that were not significant ($F(3,57) = 0.227$, $p = 0.877$).

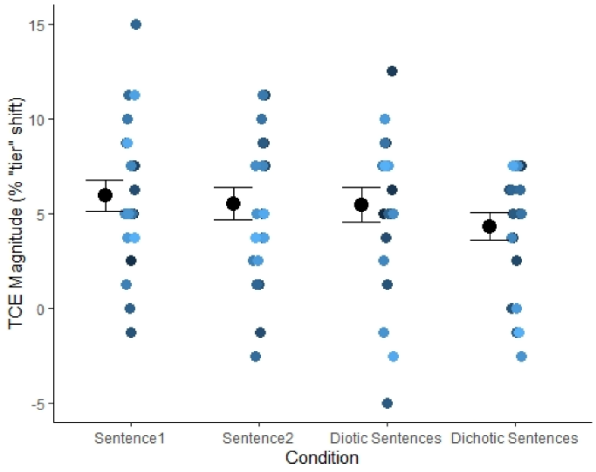

Figure 4: Experiment 2 percent "tier" shifts (TCEs) as a function of condition. Blue dots represent data from individual participants, and black circles depict condition means ± one standard error.

The mean TCEs and standard errors were calculated for each condition of Experiment 2: (Figure 4): Sentence 1 (*M* = 5.97%, *SE* = 0.8), Sentence 2 (*M* = 5.51%, *SE* = 0.8), Diotic (*M* = 5.45%, *SE* = 0.9), Dichotic (*M* = 4.32%, *SE* = 0.7). The one-way repeated measures ANOVA produced results that were also not significant ($F(3,63) = 0.806$, $p = 0.495$).

Additionally, an ANOVA across the two experiment was calculated. A mixed ANOVA was employed where Experiment (two levels; between-subjects), Condition (four levels; within-subjects), and their interaction analyzed differences in TCE magnitudes (DV) across the two experiments. TCEs did not differ by condition ($F(3,139) = 0.77$, $p = 0.51$). This was expected to occur based on the by-experiment analyses. TCEs did differ by experiment ($F(1,139) = 9.94$, $p = 0.002$). TCEs were larger in Experiment 2 (mean TCE = 5.3% shift) than Experiment 1 (mean

TCE = 3.3% shift). There was no interaction between condition and experiment ($F(3,139) = 0.13$, $p = 0.94$).

Discussion

Bosker et al. (2020) argued that selective attention didn't play a role in TCEs. However, their paradigm was not the strongest test of this question by presenting context sentences from different talkers to different ears allowing participants to separate their attention much easier. The current paradigm employed context sentences and target sounds spoken by the same talker. Also, they were presented to different ears or the same ear limiting the participants' abilities to effectively separate them.

Overall, hearing two simultaneous context sentences spoken by the same talker did not significantly change the magnitudes of TCEs in either experiment. These results corroborate Bosker's et al. (2020) findings even though variation in talkers existed. Thus, it can be concluded that rate normalization or TCEs operate independently from selective attention with or without talker variation.

TCEs are thought to be affected by qualities like speech duration and the amplitude envelope (changes in amplitude of sound over time) of speech. Amplitude envelope is a significant property of sound that allows individuals to identify sounds often with little effort. This may be the result of neurobiological mechanisms that regulate oscillatory entrainment (neuronal phase locking to specific stimulus properties such as modulation frequency). Recent studies (Giraud & Poeppel, 2012) provide some evidence that neural oscillators (theta range 3-9 Hz) regulate entrainment to the syllabic rhythms of speech. This range of frequencies overlap with most speech rates (syllables/second). Oscillatory entrainment has been viewed widely as an important concept in speech perception. Few researchers have extended this as a candidate

mechanism for TCEs (Bosker & Ghitza, 2018). However, this subtlety is important considering the neural mechanism underlying rate normalization has yet to be solidified.

There is the possibility of another mechanism responsible for rate normalization. In Oganian and Chang (2019), their focus was to target the area of the auditory cortex responsible for detecting acoustic onset edges, or rapid increases in amplitude at the beginning of a modulation. They used electrocorticography (ECoG) on neural populations of the superior temporal gyrus (STG) to study speech processing mechanisms based on natural and slowed speech. By using slowed speech, they were able to separate edges from peaks (the maximum-amplitude region of a modulation) to have them make different predictions when compared to medium-rate speech (where edges and peaks happen in rapid succession, making the same predictions). After analyzing participants' responses to slowed speech, they discovered acoustic onset edges were a better predictor for encoding amplitude envelope. This was important for understanding the linguistic structure of acoustic onset edges with their relation to vowel onsets and what cortical structure was responsible for comprehending speech at the syllabic level (Oganian & Chang, 2019).

A subsequent study by this group (Kojima et al., 2021) analyzed temporal dynamics of neural responses through magnetoencephalography (MEG) and inter-event phase coherence (IEPC) of continuous speech via evoked responses and oscillatory entrainment. The experimental paradigm focused on frequencies of the delta-theta band frequencies (1-10 Hz) produced from natural and slowed speech since these models could remain in the theta range of neural oscillators and allow acoustic onset edges of the waveform to be measured accurately (Kojima et al., 2021). Acoustic onset edges, but not amplitude envelope events, produced phase locking (specific stimuli determines how responsive a neuron is depending on its firing

frequency) of evoked responses across areas of the auditory cortex. This data corroborated the results of Oganian and Chang (2019), but again, raises the question of which mechanism is responsible for rate normalization, or TCEs.

These previous findings may reflect why the results of the current experiment proved to be statistically insignificant (TCEs were of similar magnitudes in each block). The sentences from Experiment 1 contained 10 syllables while the sentences from Experiment 2 consisted of 13. When the two sentences from the same talker were presented simultaneously, the fusion of both sentences had the perceptual effect of increasing the number of syllables heard. For instance, in Experiment 1 when the sentences were sped up, 10 syllables sounded like 11 (or a speaking rate of 10.49 syllables/sec). When the sentences were slowed down, 10 syllables sounded like 13 (or a speaking rate of 4.13 syllables/sec; shown in Figure 1). The fast sentences employed here began to fall out of the theta range which was proposed as an important factor of oscillation entrainment. In Experiment 2, when the sentences were sped up, 13 syllables sounded like 14 (or a speaking rate of 12.70 syllables/sec). When the sentences were slowed down, 13 syllables sounded like 16 (or a speaking rate of 4.84 syllables/sec; shown in Figure 2). Again, the fast sentences were outside of the theta range, but TCEs were still observed. This difference may account for why TCEs were significantly larger in Experiment 2 than in Experiment 1. Speaking rate is dependent on syllables per second, thus, affecting TCEs.

For future studies, the stimuli of this experimental paradigm could be altered by producing multiple variations in speaking rate to measure TCEs. By change speaking rate by small/medium/large amounts, one may see small/medium/large TCEs. This has been confirmed already (Summerfield, 1981). However, if one presents a fast and a slow sentence but vary the amplitude envelope to make the acoustic edges sharper/larger, this could appropriately test their

contribution to TCEs. If evoked models based on edges underlie TCEs, then one should see TCE magnitudes change as the acoustic onset edges become steeper or flatter while leaving speaking rate constant (in both the fast and slow sentences).

References

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgements. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 103-113). London: Academic Press.

Boersma, P. & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.52, retrieved 25 August 2021 from http://www.praat.org/

Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. *Language, Cognition and Neuroscience*, *33*(8), 955–967. https://doi.org/10.1080/23273798.2018.1439179

Bosker, H. R., Reinisch, E., & Sjerps, M. (2020). Temporal contrast effects in human speech perception are immune to selective attention. *Scientific Reports. 10*(1), pp. 1-11.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America. 109*(3), pp. 1101-1109. https://doi.org/10.1121/1.1345696

Kojima, K., Oganian, Y., Cai, C., Findlay, A., Chang, E., & Nagarajan, S. (2021). Low-frequency neural tracking of speech amplitude envelope reflects the convolution of evoked responses to acoustic edges, not oscillatory entrainment. *BioRxiv*, 2020-04.

Litovsky, R. Y. (2012). Spatial release from masking. *Acoustics Today*, 8(2), 18-25.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338-352.

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science Advances*, 5(11), eaay6279, 1-13.

Pardo, J. S.; Nygaard, L. C.; Remez, R. E.; Pisoni, D. B. (2021). *The handbook*

of speech perception. Second Edition. John Wiley & Sons, Inc.

Sharpe, C. M. (2021). The role of talker in adjusting for different speaking rates in speech perception. Senior Honors Thesis, University of Louisville.

Stilp, C. E. (2020). Acoustic context effects in speech perception. *WIREs Cognitive Science*, *11*(1), 1–18. https://doi.org/10.1002/wcs.1517

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074–1095. https://doi.org/10.1037/0096-1523.7.5.1074