Attention, Perception, & Psychophysics manuscript No.
(will be inserted by the editor)

Units of processing in perceptual normalization for speaking rate

Margaret Cychosz[1] · **Rochelle S. Newman**[1]

the date of receipt and acceptance should be inserted later

Abstract Because speaking rates are highly variable, listeners must use cues like phoneme or sentence duration to scale or normalize speech across different contexts. Scaling speech perception in this way allows listeners to distinguish between temporal contrasts, like voiced and voiceless stops, even at different speech speeds. It has long been assumed that this normalization or adjustment of speaking rate can occur over individual phonemes. However, phonemes are often undefined in running speech, so it is not clear that listeners can rely on them for normalization. To evaluate this, we isolate two potential processing units for speaking rate normalization—the phoneme and the syllable—by manipulating phoneme duration in order to cue speaking rate, while also holding syllable duration constant. In doing so, we show that changing the duration of phonemes both with unique acoustic signatures (/kɑ/) and overlapping acoustic signatures (/wɪ/) results in

This work was supported by National Institute on Deafness and Other Communication Disorders grants T32DC000046 and F32DC019539 (M.C.) and 5R01HD081127 (R.S.N.).

[1]Department of Hearing & Speech Sciences
University of Maryland, College Park
0100 Samuel J. LeFrak Hall
College Park, MD, 20742
E-mail: mcychosz@umd.edu

a speaking rate normalization effect. These results suggest that even absent clear acoustic boundaries within syllables, listeners can normalize for rate differences on the basis of individual phonemes.

Keywords speaking rate · speech perception · normalization · speech processing · categorical perception

1 Introduction

Speaking rate varies widely between and within speakers. Yet many phonological contrasts of the world's languages rely on temporal cues, such as vowel length or voice onset time. Consequently, listeners must perceptually *normalize* for speaking rate, remapping acoustic cues across different contexts and speakers, in order to comprehend speech and acquire language.

Listeners employ perceptual normalization (or compensation) for speaking rate over a variety of levels in the speech signal.[1] For example, to categorize a temporally-cued contrast like /k-g/, listeners could use *proximal* information in the speech signal, like the duration of vowels or consonants that are adjacent to the target phoneme (Diehl and Walsh 1989; Miller and Liberman 1979; Newman and Sawusch 1996; Summerfield 1981). Listeners could also use *distal* information in the speech signal like the rate of the overall sentential context, another talker's habitual or situational speaking rate, or even the duration of non-speech stimuli like tones (Maslowski et al. 2019; Newman and Sawusch 2009; Reinisch 2016; Wade and Holt 2005). In both cases, for a contrast like /k-g/, shorter-duration cues (e.g., a shorter adjacent consonant or a faster sentence) suggest a faster speaking rate and therefore bias listeners to the positive voice onset time phoneme /k/. In contrast, longer-duration cues bias listeners to the negative or neutral voice onset time phoneme /g/.

[1] Throughout the paper, we refer to normalization for speaking rate without implying that listeners normalize for all contextual information during perception. We also do not use the term "normalization" to imply that listeners eliminate vs. maintain rate-based information.

Research on proximal information for speaking rate normalization has focused on cues such as the duration of phones preceding or following the target segment. As a result, we now know that although there are more degrees of freedom in vowel than consonant duration (Crystal and House 1988), both vowels and consonants can provide rate normalization cues (Diehl and Walsh 1989; Summerfield 1981; see Toscano and McMurray 2012 for an alternative interpretation). Additionally, while more distal cues for speaking rate normalization, like preceding word duration, are well-documented (see previous citations), there is some evidence for an adjacency bias in speaking rate normalization: listeners may only scale for speaking rate over a limited temporal window of a single adjacent phoneme or syllable (Newman and Sawusch 1996; Sawusch and Newman 2000; cf. Baese-Berk et al. 2014).

While careful experimental manipulations have led us to understand *which* cues listeners use during perceptual normalization for speaking rate, less is known about *how* the cues are incorporated. This gap in our understanding of rate normalization processes is relevant for a number of theoretical and practical reasons. Research into proximal cues for rate normalization has traditionally assumed that phonemes are the basic unit over which speaking rate can be normalized. But this assumption may be premature. For one thing, it is difficult for listeners to isolate phonemes in the comprehension of spontaneous, running speech. Articulatory undershoot and hypoarticulation reduce phonological contrasts (Johnson et al. 1993; Lindblom 1990). Coarticulation blurs acoustic boundaries between adjacent phones as speakers consistently anticipate upcoming speech sounds (Whalen 1990). Some phoneme pairs, like glides and vowels or laterals and vowels, are especially susceptible to these coarticulatory pressures and less able to resist the influence of adjacent segments (Recasens 1985). In all, phonemes, particularly voiced, non-strident phonemes, are not reliably discriminable. It is therefore plausible that listeners would instead normalize speaking rates over syllables or other other acoustically-based segments which may not straightforwardly correspond to linguistic representations.

Rate normalization has often been considered a low-level, domain-general auditory process (Bosker 2017): it is involuntarily activated after milliseconds of exposure to a speech- or non-speech-like stimulus (Reinisch 2016) and has been documented in non-human (avian) species (Welch et al. 2009). However, it is also increasingly apparent that several higher-level constructs such as language experience (Baese-Berk et al. 2016), listener familiarity with the speaker (Kleinschmidt 2016; Reinisch 2016), and some aspects of language structure such as intonation (Steffman 2019) also mediate rate normalization. It is thus possible that rate normalization interacts with additional higher-level linguistic units, such as the syllable, although this has not been empirically tested.

From a machine learning perspective, invariance in the speech signal is a central obstacle to achieving higher-performing speech-to-text and automatic speech recognition applications. Understanding appropriate mechanisms for normalization, including rate normalization, in human listeners may facilitate machine performance, as it may be simpler to program normalization on the basis of acoustically-driven signals (such as syllables) than linguistically-driven concepts (phonemes). If human listeners reliably normalize for speaking rate at the phonemic level, even in the absence of explicit acoustic signals, it would suggest that phonemic structure should be incorporated into natural language processing algorithms to benefit machines' learning of spontaneous speech.

The present experiments were designed to investigate the effects of acoustic separability, or the ability to distinguish between two adjacent phonemes, on speaking rate normalization. The overarching goal is to understand the processing units (syllable or phoneme) involved in the perceptual normalization of speaking rate. In a pair of phoneme category rating experiments, we asked whether two syllables containing phones differing in acoustic separability (acoustically-distinct /kɑ/ versus overlapping /wɪ/) would result in separate rate normalization effects or in a single combined rate normalization effect. We chose to evaluate the effects of speaking rate upon the perception of the /ʃ-tʃ/ contrast in American English as this con-

trast has demonstrated a rate normalization effect in prior research (Newman and Sawusch 1996). In that work, the authors were able to trigger a /ʃ-tʃ/ phonetic boundary shift in a nonce word series ranging from /ʃkɑs/ - /tʃkɑs/ ("shkas" to "chkas") by adjusting the duration of /k/ in the stimuli. Ambiguous stimuli, with a longer /k/ duration, suggested a slower speaking rate and biased listeners to perceive /tʃ/ while a shorter /k/ suggested a faster speaking rate and biased listeners to perceive /ʃ/.

A limitation of Newman and Sawusch (1996) and other previous work on this topic is that changes to the duration of a single phoneme, like /k/, also rendered changes to the duration of the surrounding syllable (e.g. /kɑ/) and word (e.g. /ʃkɑs/): a longer-duration /k/ resulted in a longer /kɑ/ syllable and /ʃkɑs/ word. As a result, any rate normalization effect could just as easily be attributed to the duration of the manipulated phoneme as the duration of the entire syllable or word.

To isolate phonemes as the potential processing unit in speaking rate normalization, Experiment 1 uses the same /ʃkɑs/-/tʃkɑs/ series as previous work but varies the syllable nucleus /ɑ/ duration in the opposite direction of /k/. This manipulation leads to a /ʃkɑs/-/tʃkɑs/ series with consistent syllable and word, but varying phoneme, durations. Any rate normalization effect for these stimuli thus cannot be due to syllable or word durations, as the series did not differ in these respects. Instead, the normalization effect could only be caused by variation in the manipulated phoneme /k/. Finding a rate normalization effect would suggest that the /k/ was treated as a separate unit from the following vowel and that rate normalization took place over phoneme-sized units.

Varying the nucleus duration in the opposite direction of the consonant is unlikely to cancel out any potential effect of the consonant's duration because duration effects are (1) weighted by distance and /k/ is linearly closer to the target contrast in the /ʃkɑs/-/tʃkas/ series and (2) proportional and /k/ is much shorter than /ɑ/ so similar durational changes (e.g. 20 ms) have disproportionate impacts

upon /k/ and /ɑ/. Indeed, we do find a rate normalization effect in Experiment 1, suggesting that changing the vowel duration in the opposite direction did *not* cancel out any consonant duration effect. Finding a rate normalization effect in the Experiment 1 stimuli leads us to conduct Experiment 2 where we again test for rate normalization effects but using syllables that contain less discriminable phones. We use a similar nonce word series ranging from /ʃwɪb/ - /tʃwɪb/ ("shwihb" to "chwihb") where we manipulated the duration of /w/ in /wɪ/. Although we did find an effect of /k/ duration upon perception of the initial /ʃ-tʃ/ contrast in Experiment 1, suggesting phoneme-level processing during rate normalization, we hypothesized that we may *not* find this same effect of /w/ duration on the same /ʃ-tʃ/ contrast in Experiment 2, suggesting higher-level (syllable or word) processing for less-discriminable phones.

2 Experiment 1

2.1 Methods

Participants Twenty-nine members of the University of Iowa community participated in this experiment for course credit. All listeners were native speakers of American English, and had no reported history of a speech or hearing impairment. In later questioning, one of these participants was found not to be a native English speaker; his data were not examined. Seven of the listeners failed to respond on at least 80% of the trials. It is unclear why these participants failed to respond on all trials; perhaps they simply failed to press the buttons on the response box firmly enough for the computer to register their decision. However, as these participants only provided responses on a small portion of the trials, far too few to impute the missing data, their data were likewise eliminated. This resulted in a total of 21 listeners.

Stimuli An adult native English-speaking man was recorded producing the syllable /ʃkɑs/ in a fluent speech context. His speech was amplified, low-pass filtered

at a 4.8 kHz sampling rate, and digitized via a 12-bit, analog-to-digital converter at a 10-kHz sampling rate. The initial consonant /ʃ/ was then separated from the remainder of the syllable, with the boundary being the onset of closure for the following /k/. A continuum of ten items, /ʃ/-/tʃ/, was then created by removing successive 10-ms sections from the /ʃ/ onset. A linear amplitude ramp, with duration varying along with frication duration, was used over the initial portion of each token to give the items a more natural attack. The duration of the ramp varied from 6 to 60 ms, with a 9 ms step. The resulting series ranged from 60-150 ms in duration, with the longer frication sounding more similar to a /ʃ/ and the shorter frication sounding more similar to a /tʃ/. Further details on the original stimulus creation can be found in Newman and Sawusch (1996).

The remainder of the word—the syllable /kɑs/—was edited to create two new syllables, one with a shorter /k/ (and longer /ɑ/) and one with a longer /k/ (and shorter /ɑ/). We interpreted the /k/ to include the closure, burst, aspiration, and first four pitch pulses (which appeared to correspond to the transition of the first formant). The duration of this base /k/ was approximately 1/3 that of the vowel (see Figure 1). Thus, an equivalent amount of change in duration for /k/ and /ɑ/ will be much larger proportionately for /k/.

The duration of /k/ was altered by removing or reduplicating pitch pulses and sections of burst and aspiration. Only short, nonadjacent sections of burst and aspiration were deleted or reduplicated so as to maintain the general amplitude profile and prevent the perception of frozen noise. No change was made to the closure duration; although closures do tend to vary slightly with speaking rate, this variability is typically quite small (Crystal and House 1988; Gay 1978), and thus unlikely to have a substantial perceptual effect. For the short /k/ stimulus, two pitch pulses were removed, as well as 17.2 ms of the burst and aspiration; for the long /k/ stimulus, four pitch pulses and 22 ms of the burst and aspiration were reduplicated. The vowel duration was similarly adjusted by removing or reduplicating nonadjacent pitch pulses, so as to make the absolute amount of change in

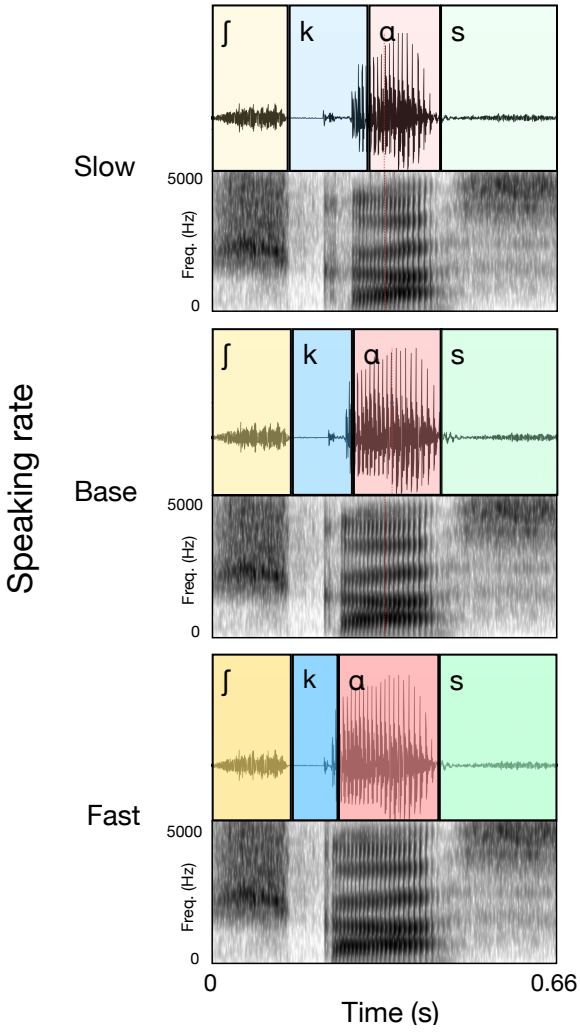

Fig. 1
Speaking rate manipulations and stimuli duration for first step of series: Experiment 1.

the vowel as close as possible to the absolute amount of change in the stop consonant. The original /k/ stimulus served as the intermediate duration stimulus resulting in a 3-way /k/-duration series (short/fast, baseline/intermediate, and long/slow), although we make no claims as to it actually being half-way between the other two stimuli perceptually. The short /k/, base /k/, and long /k/ versions of the syllable were then appended to each member of the 10-item /ʃ/-/tʃ/ series.

Procedure Participants completed 1 practice/training block of 60 trials and 4 test blocks of 90 trials each. Responses from the training block were not analyzed. The four 90-trial test blocks were comprised of three repetitions of each of the 30 stimuli (3 /k/ durations X 10-step /ʃ-tʃ/ continuum) for a total of 360 trials per participant, or 12 repetitions of each stimulus.

Stimuli within each block were presented randomly to listeners using a Macintosh 7100/AV computer at a comfortable listening level over Audiotechnica ATH-M40 headphones. Listeners were prompted with each stimulus and asked to rate the quality of the initial phoneme on a six-point scale, ranging from "a good *sh*" to "a good *ch*", by pressing the appropriate button on a computer-controlled response box. The use of ratings allowed the detection of subtle differences within a category that may not have been obvious with simpler, categorical labeling (Sawusch 1976). Presentation pace depended on the subject's response rate. Each trial began 1000 ms after the listener had responded to the previous trial, or after an interval of 3000 ms following stimulus onset, whichever came first. The experiment lasted approximately 45 minutes.

2.2 Results

Data were analyzed in the RStudio computing environment (version: 1.4.1103; RStudioTeam 2020). Visualizations were created with `ggplot2` (Wickham 2016). Modeling was conducted and presented using the `lme4` (Bates et al. 2015), `lmerTest` (Kuznetsova et al. 2017), and `broom.mixed` (Bolker and Robinson 2020) packages. Model parameter significance was determined via a combination of log-likelihood comparisons between models, AIC estimations, and p-values from model summaries.

To test for an effect of phoneme duration on rate normalization, we modeled two different outcome variables: percentage of /ʃ/ responses and /ʃ/-ness ratings. For the percentage of /ʃ/ responses, an average /ʃ/ response was calculated for each participant, for each stimulus item (Figure 2), while /ʃ/-ness ratings were

simply computed for each individual stimulus item presented (item-level effect) (Figure 3). We elected to model two outcomes because traditional work on rate normalization has modeled percentage phoneme responses grouped over stimuli repetitions (e.g. Diehl and Walsh 1989), while newer work has been able to model item-level effects (e.g. Maslowski et al. 2019) and we wished to make our work comparable to both of these domains.

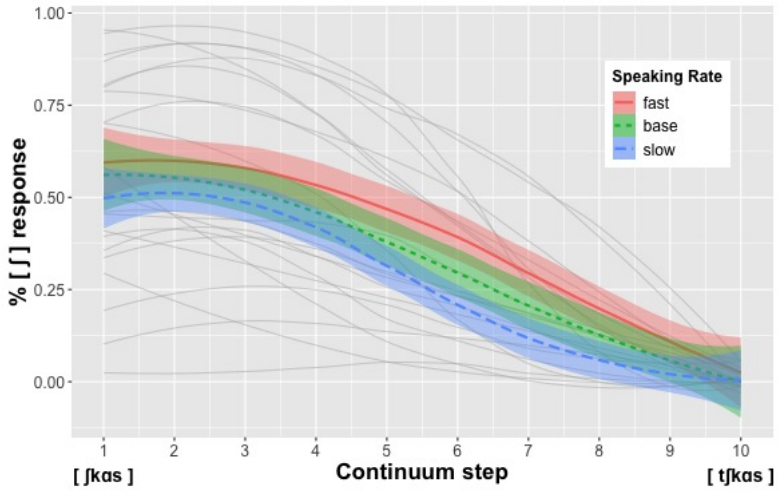

Fig. 2 Spaghetti plot of percentage /ʃ/ response by series step and speaking rate: /k/ duration manipulation. Thick, darker lines represent group averages by speaking rate and lighter lines represent individual participant responses. Ribbons represent 95% confidence intervals.

Figures 2 and 3 suggest the presence of a rate normalization effect from phoneme duration manipulations. The confidence intervals surrounding the speaking rate conditions (Slow, Base, Fast) do not overlap in the middle, ambiguous section of the continuum. More specifically, we see the effect in the expected direction: slower speaking rates bias /tʃ/ responses and higher /tʃ/ ratings, while faster rates bias /ʃ/ responses and higher /ʃ/ ratings.

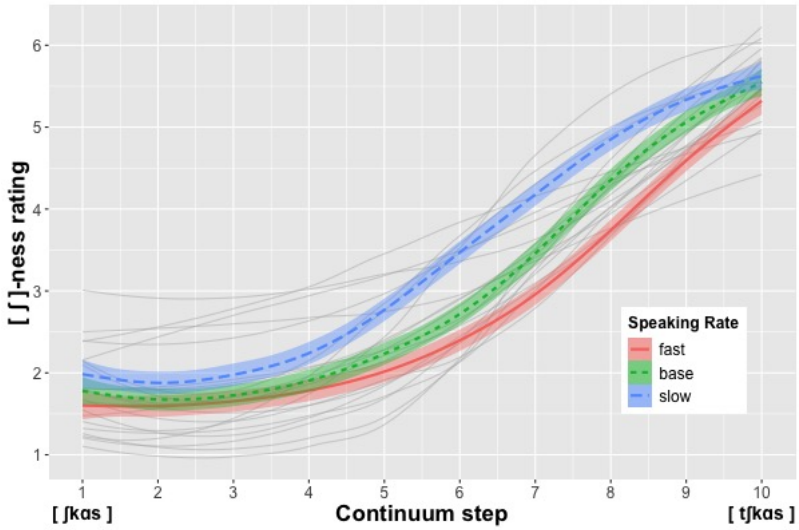

Fig. 3 Spaghetti plot of /ʃ/-ness ratings (1=good /ʃ/, 6=good /tʃ/) by series step and speaking rate: /k/ duration manipulation. Thick, darker lines represent group averages by speaking rate and lighter lines represent individual participant responses. Ribbons represent 95% confidence intervals.

To further examine a potential rate normalization effect, we fit models to our two outcome variables. Both models were fit to ambiguous items in the middle of the stimuli series, at the categorical perception boundary (steps 4-7 in this continuum, following Bidelman et al. (2019) who modeled steps 3-5 on a 7-point continuum). To predict the percentage of /ʃ/ responses, we fit a linear mixed effects model with the maximal random effect structure that permitted model convergence. This model included random slopes of Speaking Rate by Participant. Slope terms for interactions of Speaking Rate and Continuum Step did not converge, nor did models with random intercepts of Participant together with slopes of Speaking Rate by Participant. The effect of Speaking Rate (modeled categorically as "Slow," "Base," and "Fast") improved upon the random effects-only model as did Continuum Step (modeled as a continuous variable) (Table 1).[2] Unsurprisingly, the percentage of /ʃ/ responses decreased with increased steps along the continuum (β=-0.09, z=-12.15, p<.001). For Speaking Rate, there was a higher percentage

[2] In the random effects, Continuum Step was modeled as three-way categorical variable; these models did not converge.

Table 1 Model predicting percentage /ʃ/ response: Experiment 1

Parameter	Estimate	S.E.	z-statistic	p-value	95% CI
Intercept	0.56	0.06	9.86	<.001	0.45 - 0.67
Rate:Fast	0.09	0.02	3.72	0.001	0.04 - 0.14
Rate:Slow	-0.07	0.03	-2.54	0.019	-0.12 - -0.02
Continuum Step	-0.09	0.01	-12.15	<.001	-0.1 - -0.07

Table 2 Model predicting /ʃ/-ness ratings: Experiment 1

Parameter	Estimate	S.E.	z-statistic	p-value	95% CI
Intercept	1.31	0.18	7.16	<.001	0.95 - 1.66
Rate:Fast	-0.27	0.13	-2.14	0.052	-0.52 - -0.02
Rate:Slow	0.59	0.13	4.67	0.005	0.34 - 0.84
Continuum Step	0.50	0.05	10.79	<.001	0.41 - 0.59

of /ʃ/ responses in the Fast condition than the Base condition (β=0.09, z=3.72, p=0.001) and a lower percentage of /ʃ/ responses in the Slow condition than Base (β=-0.07, z=-2.54, p=0.019), suggesting a rate normalization effect.

To model item-level effects, we fit a second model to predict /ʃ/ ratings (1-6 scale where a lower rating indicates more /ʃ/-ness and a higher rating indicates more /tʃ/-ness). The random effect structure again included the maximal number of terms that permitted model convergence, in this case random effects of Participant and Item. There were significant main effects of Continuum Step and Speaking Rate (Table 2). Ratings increased with continuum steps, indicating increased perception of /tʃ/ (β=0.50, z=10.79, p<.001). Most importantly, the Slow speaking rate condition predicted higher ratings, or more /tʃ/-ness, than the Base speaking rate condition (β=0.59, z=4.67, p=.005): a longer /k/ duration, suggesting a slower speaking rate, biased listeners to perceive and rate the stimuli as more /tʃ/-like. However, there was only a reliable effect of Speaking Rate in the Slow condition; the difference between the Fast and Base speaking rates upon listeners' rankings approached but did not reach significance.

Overall, these results demonstrate that manipulating /k/ duration, while holding the syllable duration constant, significantly affected the percentage of /ʃ/ re-

sponses and /ʃ/-ness ratings, especially in the Slow speaking rate condition, suggesting that listeners can normalize for speaking rate over individual phonemes.

2.3 Interim discussion

Experiment 1 demonstrated that two phonemes with obvious acoustic boundaries, /k/ and /ɑ/, were treated as separate units during rate normalization. This result implies that the processing unit during rate normalization is something smaller than a syllable. However, /k/ and /ɑ/ are fairly acoustically distinct and easy to distinguish from one another. Do listeners scale for speaking rate over phoneme-sized units that are difficult to distinguish? Experiment 2 examines a syllable containing phonemes that are much more difficult to segment acoustically: a glide and a vowel. To examine this, we chose a nonce word series that ranged from /ʃwɪb/-/tʃwɪb/. Previous work on similar stimuli—a /swæb/-/twæb/ continuum—demonstrated that varying the /w/ duration while leaving the vowel constant, and varying the /æ/ duration while leaving the glide constant, both lead to a change in category boundary location for the initial /s-t/ contrast (Newman and Sawusch 1996). Yet the effect could have been driven by the duration of a unit larger than the phoneme, because changing the /w/ duration while leaving the /æ/ constant results in the combined syllabic unit also being longer. However here, as in Experiment 1, we varied the /w/ duration while also altering the /ɪ/ duration in the opposite direction, leading once again to a series with consistent syllable and word durations. If /w/ and /ɪ/ are treated as separate units during rate normalization, as /k/ and /ɑ/ were, then manipulating the duration of /w/ should lead to a rate normalization effect in this series.

3 Experiment 2

3.1 Methods

Participants Twenty-two members of the University of Iowa community participated in this experiment for course credit. All were native English speakers with no reported history of a speech or hearing impairment. Three participants did not respond on at least 80% of the trials, so their data were removed from analysis leaving 19 participants.

Stimuli Stimulus creation was nearly identical to that in Experiment 1. The same speaker produced the syllable /ʃwɪb/ in the same manner previously described. The initial fricative was separated from the remainder of the syllable, with the boundary being the zero-crossing preceding the first pitch pulse of the /w/. A series of ten items ranging from /ʃ/ to /tʃ/ was created in a similar manner as Experiment 1, by removing successive sections of approximately 10 ms from the onset of the /ʃ/.

The syllable /wɪ/ was edited in the same manner as the /kɑ/ syllable in Experiment 1. Based on spectral analysis, the first 7 vocal pulses were considered part of /w/ rather than the /ɪ/, because these pulses appeared to constitute the /w/ formant transitions (especially those of the first formant). We lengthened and shortened the /w/ and /ɪ/ durations by reduplicating or deleting nonadjacent pitch pulses in the same manner as before. For the shorter /w/, three pitch pulses were removed, whereas four pulses were reduplicated to create the long /w/ (and pitch pules from the vowel were likewise removed or reduplicated in the same manner to keep the syllable duration constant). The original items served as the intermediate duration. The /w/ duration was shorter than that of the /ɪ/, so the same amount of absolute change resulted in a larger change proportionately for the /w/ than for the vowel. The short /w/, baseline/intermediate /w/, and long /w/ versions of the syllable were then appended to each member of the 10-item /ʃ/-/tʃ/ series. This

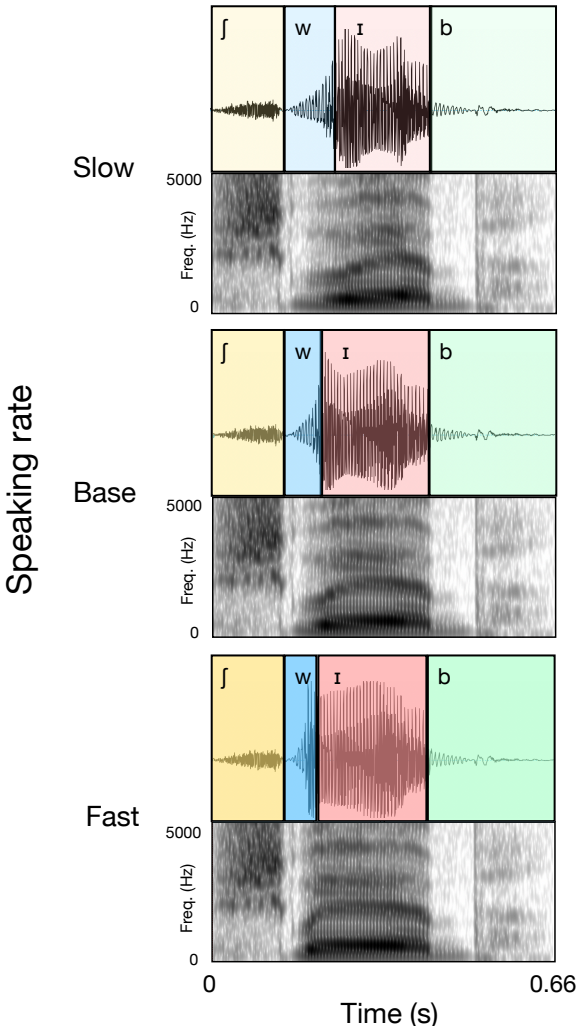

Fig. 4
Speaking rate manipulations and stimuli duration for first step of series: Experiment 2.

resulted in three /w/-duration series with a constant syllable and word duration, but varying /w/ (and vowel) durations.

Procedure The procedure was identical to that of Experiment 1.

3.2 Results

As in Experiment 1, to evaluate a potential rate normalization effect we modeled two different outcome variables: percentage of /ʃ/ responses and /ʃ/-ness ratings. Again, an average /ʃ/ response was calculated for each participant (Figure 5) and /ʃ/-ness ratings were computed for each individual stimulus (Figure 6). The visualizations suggest an effect of speaking rate (/w/ duration) upon /ʃ/ responses and /ʃ/ ratings in the same direction as Experiment 1: slower speaking rates bias more /tʃ/ responses.

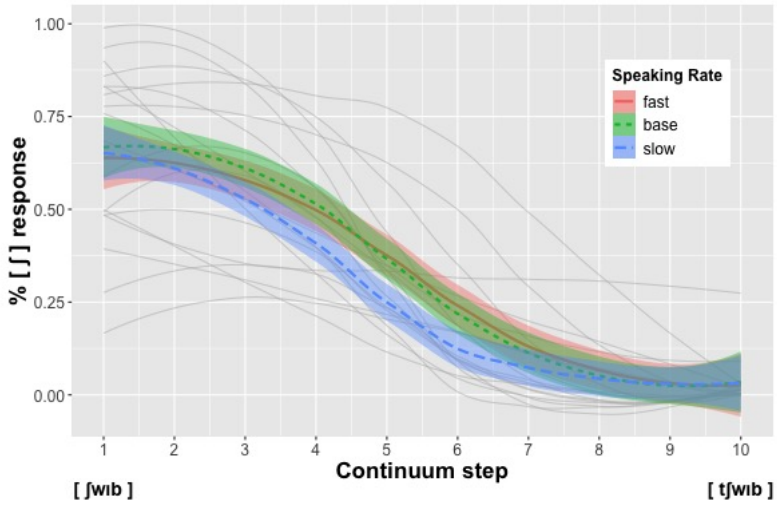

Fig. 5 Spaghetti plot of percentage /ʃ/ response by series step and speaking rate: /w/ duration manipulation. Thick, darker lines represent group averages by speaking rate and lighter lines represent individual participant responses. Ribbons represent 95% confidence intervals.

As before, we fit a series of models to the ambiguous items in the middle of the stimuli series, at the categorical perception boundary (steps 4 through 7). To predict the percentage of /ʃ/ responses, we fit a linear mixed effects model that included random slopes of Speaking Rate by Participant. Like Experiment 1, there were significant main effects of Speaking Rate and Continuum Step: the percentage of /ʃ/ responses increased as the continuum step increased (more /tʃ/-like stimuli) (β=-0.13, z=-14.44, p<.001). There was a significantly smaller percentage of /ʃ/

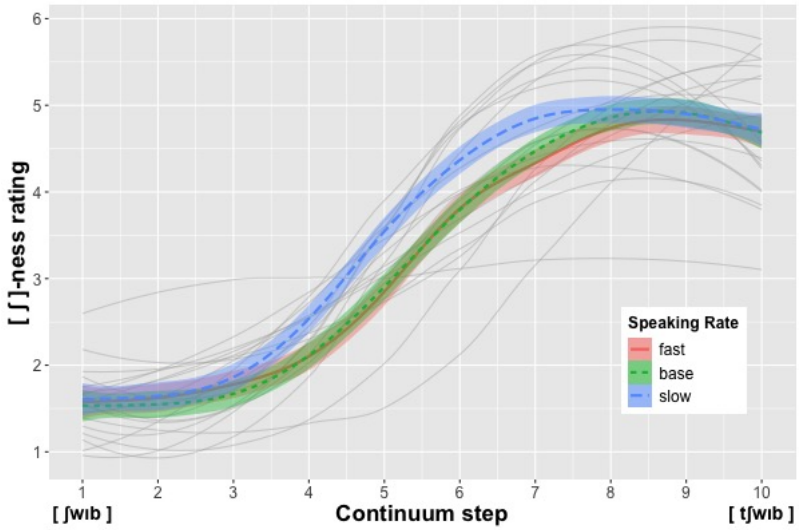

Fig. 6 Spaghetti plot of /ʃ/-ness ratings (1=good /ʃ/, 6=good /tʃ/) by series step and speaking rate: /w/ duration manipulation. Thick, darker lines represent group averages by speaking rate and lighter lines represent individual participant responses. Ribbons represent 95% confidence intervals.

Table 3 Model predicting percentage /ʃ/ response: Experiment 1

Parameter	Estimate	S.E.	z-statistic	p-value	95% CI
Intercept	1.00	0.07	15.33	<.001	0.88 - 1.13
Rate:Fast	0.02	0.02	0.69	0.49	-0.03 - 0.06
Rate:Slow	-0.09	0.03	-2.66	0.02	-0.16 - -0.02
Continuum Step	-0.13	0.01	-14.44	<.001	-0.15 - -0.11

responses in the Slow speaking rate condition than Base condition (β=-0.09, z=-2.66, p=0.02), demonstrating a rate normalization effect for these stimuli, but no difference in the percentage of /ʃ/ responses between the Fast speaking rate condition and the Base condition (Table 3).

Finally, we fit a model to predict /ʃ/-ness ratings for these stimuli, where a lower rating indicated that listeners considered the stimuli more /ʃ/-like. The random effect structure again included random intercepts by Participant and Item and there were significant main effects of Continuum Step and Speaking Rate (Table 4). Unsurprisingly, stimuli further along the /ʃ-tʃ/ continuum were perceived as more /tʃ/-like (β=0.82, z=17.66, p<.001). Longer /w/ durations, indicating a

Table 4 Model predicting /ʃ/-ness ratings: Experiment 2

Parameter	Estimate	S.E.	z-statistic	p-value	95% CI
Intercept	-1.19	0.30	-3.99	0.002	-1.78 - -0.61
Rate:Fast	-0.06	0.13	-0.49	0.64	-0.31 - 0.19
Rate:Slow	0.54	0.13	4.20	0.003	0.29 - 0.79
Continuum Step	0.82	0.05	17.66	<.001	0.73 - 0.91

slower speaking rate, also biased listeners to perceive the stimuli as more /tʃ/-like (β=0.54, z=4.20, p=.003). As with Experiment 1, we did not find an effect of speaking rate in the Fast speaking rate condition. Overall, the results from Experiment 2 show an effect of speaking rate (/w/ duration) upon the perceived phonetic boundary between /ʃ/ and /tʃ/, indicating that normalization can occur over these phonemes without clear acoustic boundaries.

4 General discussion

Listeners must compensate for variation across different speakers, in different contexts, to comprehend speech and language. Normalization for speaking rate is one important example of this process: it allows listeners to maintain temporal contrasts, such as voice onset time, across different speech speeds and between different speakers. In a pair of experiments, we evaluated whether listeners could use information from individual phonemes—which coarticulation and hypoarticulation often render undefined in the acoustic signal—instead of syllables to normalize for speaking rate. Listeners did normalize over phonemes, including acoustically-overlapping phonemes, to factor out speaking rate, demonstrating that the phoneme is a potential processing unit for rate normalization processes.

Work on proximal information in the speech signal for rate normalization has long suggested that normalization occurs over individual phones (Diehl and Walsh 1989; Newman and Sawusch 1996). Empirical support was lacking, however, because previous work altered the duration of the carrier syllable and word in addition to the phone. Here we compensated for changes in consonant duration by

also changing the nucleus duration. This step allowed us to maintain a consistent syllable duration, avoid the previous experimental confound, and isolate effects of phoneme duration on rate normalization. Since we replicated previous work in finding an effect of phoneme duration on this phonetic boundary shift, we can now more definitively say that listeners can use phonemes to compute speaking rate. Furthermore, by also evaluating the effects of acoustic distinctiveness on rate normalization, we were additionally able to show that this phoneme processing for rate normalization even occurs in less-than-ideal acoustic environments.

Rate normalization can be activated after just milliseconds of exposure (Reinisch 2016), and is documented in human and non-human species alike (Welch et al. 2009), suggesting that this type of normalization is a low-level auditory process that could be partially domain-general. Finding that listeners can compute speaking rate over individual phonemes speaks directly to this idea. Phonemes do not relay a clear acoustic signal. They are indistinct, coarticulated, and reduced—traits that are exacerbated when the features (voicing, stridency) of adjacent phones overlap within syllables. If rate normalization were exclusively or primarily domain-general, it is unclear how listeners could normalize over individual phonemes. It is possible that listeners may *prefer* or *tend* to normalize over syllables, or relatively more acoustically-reliable components of speech such as word boundaries, but will compute over phonemes in the absence of higher-level information. Our experiments were not designed to contrast listeners' preferred processing unit for rate normalization. It is also possible, as Bosker (2017) suggests, that perceptual normalization for speaking rate could be domain general for some lower-level constructs, such as phonetic boundary shifts, but increasingly language-specific at higher levels such as determining the presence of function words (Dilley and Pitt 2010). Nevertheless, the fact that listeners could normalize over phonemes in these experiments is strong evidence that rate normalization processes are driven by experience with a language, instead of the raw acoustic signal alone.

The results of these experiments open up several avenues for future research. First, these experiments only tested American English listeners listening to mostly singleton consonants and monophthongal vowels embedded in nonce words. But other works have found clear effects of language structure and experience on rate normalization (Baese-Berk et al. 2016; Steffman 2019). Do listeners also normalize over units, like morae, geminates, or diphthongs that are heavier/larger than phonemes but smaller than syllables? Phonotactic structure is another unexplored aspect of language structure that may be relevant for understanding how listeners calculate speaking rate. Some languages, such as Japanese, tend to have more acoustically "confusable" internal syllable structures, only permitting nasal consonants, and not stops, in coda position for example. If the acoustic signature within syllables tends to be more indistinct, listeners could learn to rely less on individual phonemes for normalization. Studying how perceptual normalization for speaking rate develops in children would be another way to evaluate this idea.

It will also be important for future work to evaluate processing units for normalization in faster and more naturalistic stimuli as perceptual normalization for speaking rate is likely idiosyncratic and dependent upon the context and speaker (Goldinger and Azuma 2003). And more naturalistic stimuli, that contain multiple, co-varying phonetic cues (i.e formant transition duration and frequency), have previously been shown to mitigate rate normalization effects (Shinn et al. 1985). Here we originally hypothesized that listeners would normalize over syllables or other supra-phonemic chunks because phonemes are highly confusable and indistinct during comprehension especially in fast, running speech. And while our experiments instead showed reliable effects of phoneme duration on the phonetic boundary shift, the experimental stimuli clearly differed from what listeners would hear and process in real-world contexts. For example, even the manipulated consonant in the "fast" speaking rate condition in Experiment 1 was relatively slow (91 ms) compared to the word-medial stop consonants that listeners might hear in everyday conversation. For extremely fast speech, listeners might rely less

on individual phones and more on syllables or words. Faster, naturalistic speech also drives acoustic reduction and heightened coarticulation (Fourakis 1991; Gay 1981). However, these acoustic cues did not necessarily accompany the stimuli employed in these experiments as we wanted to isolate the effects of speaking rate. But extreme reduction in other, more naturalistic listening conditions could lead listeners to normalize over different units.

4.1 On cue integration versus normalization

Some recent work on proximal cues for phoneme classification has suggested that listeners may not normalize for speaking rate but rather *integrate* acoustic cues that overlap with speaking rate to classify phonemes (Toscano and McMurray 2010, 2012). For example, for voice onset time contrasts, vowel duration indicates both speaking rate and voicing: the burstiness of voiceless stops can cause the onset of the following vowel to de-voice slightly, leading it to be perceived as shorter in duration.

While it was not the goal of this study to contrast cue integration and rate normalization accounts to explain proximal effects upon phonetic boundary shifts—and the results of Toscano and McMurray (2012) do convincingly demonstrate that vowel length integration, not normalization for speaking rate, explains proximal effects upon stop voicing classifications—we believe the current results show rate effects and not the more straightforward acoustic cue integration. This is because our target contrast for both studies, /ʃ-tʃ/, was cued by the duration of the following *consonant*, not vowel (/k/ in /ʃkɑs/-/tʃkɑs/ for Exp. 1 and /w/ in /ʃwɪb/-/tʃwɪb/ for Exp. 2). But more importantly, there is no evidence that stop or glide duration reliably indicates fricative-affricate classification. And unlike the effect of stop aspiration upon perceived vowel length (aspiration causes vowel devoicing), there is no phonetic reason to assume that fricatives and affricates would have different effects on /k/ or /w/ duration or voicing. Consequently, we believe

that the effects of consonant duration upon the phonetic boundary shift between /ʃ/ and /tʃ/ in these studies indicates rate normalization, not cue integration.

5 Conclusion

Unlike previous work studying proximal effects on rate normalization, this study manipulated speaking rate via phoneme duration while holding the duration of carrier syllables and words constant. We still demonstrated rate effects upon the phonetic boundary shift between /ʃ/ and /tʃ/, both for syllables containing acoustically-distinct /kɑ/ and -overlapping phonemes /wɪ/. These results present evidence that listeners process speaking rate over individual phonemes, even in the absence of clear acoustic boundaries between phones, suggesting roles of linguistic structure and language experience for perceptual normalization of speaking rate.

Acknowledgements The authors wish to thank Jessica Burnham, Jim Sawusch, and Jan Edwards for their assistance with this work.

Data availability

Analysis scripts to replicate modeling results are included in the affiliated GitHub repository (`https://github.com/megseekosh/rate-normalization`. Experiments were not pre-registered.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

Baese-Berk, M., Morrill, T., and Dilley, L. (2016). Do non-native speakers use context speaking rate in spoken word recognition? In *Proceedings of Speech Prosody*, volume 8, pages 979–983.

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., and McAuley, J. D. (2014). Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science*, 25(8):1546–1553.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bidelman, G. M., Sigley, L., and Lewis, G. A. (2019). Acoustic noise and vision differentially warp the auditory categorization of speech. *The Journal of the Acoustical Society of America*, 146(1):60–70.

Bolker, B. and Robinson, D. (2020). Broom.mixed: Tidying methods for mixed models.

Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. *Attention, Perception, & Psychophysics*, 79(1):333–343.

Crystal, T. H. and House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *The Journal of the Acoustical Society of America*, 83(4):1553–1573.

Diehl, R. L. and Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *The Journal of the Acoustical Society of America*, 85(5):2154–2164.

Dilley, L. C. and Pitt, M. A. (2010). Altering Context Speech Rate Can Cause Words to Appear or Disappear. *Psychological Science*, 21(11):1664–1670.

Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. *The Journal of the Acoustical Society of America*, 90(4):1816–1827.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *The Journal of the Acoustical Society of America*, 63(1):223–230.

Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica*, 38:148–158.

Goldinger, S. D. and Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, 31(3-4):305–320.

Johnson, K., Flemming, E., and Wright, R. (1993). The Hyperspace Effect: Phonetic Targets Are Hyperarticulated. *Language*, 69(3):505–528.

Kleinschmidt, D. F. (2016). *Perception in a Variable but Structured World: The Case of Speech Perception*. Unpublished doctoral dissertation, University of Rochester, Rochester, NY.

Kuznetsova, A., Brockhoff, P., and Christensen, R. (2017). lmerTest Package: Tests in linear mixed-effects models. *Journal of Statistical Software*, 82(13):1–26.

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 403–439. Springer Netherlands, Dordrecht.

Maslowski, M., Meyer, A. S., and Bosker, H. R. (2019). How the tracking of habitual rate influences speech perception. *Journal of Experimental Psychology: Learning, Memory,*

and Cognition, 45(1):128–138.

Miller, J. L. and Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, 25(6):457–465.

Newman, R. S. and Sawusch, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics*, 58(4):540–560.

Newman, R. S. and Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. *Journal of Phonetics*, 37(1):46–65.

Recasens, D. (1985). Coarticulatory patterns and degrees of coarticulatory resistance in Catalan CV sequences. *Language and Speech*, 28(2):97–114.

Reinisch, E. (2016). Speaker-specific processing and local context information: The case of speaking rate. *Applied Psycholinguistics*, 37(6):1397–1415.

RStudioTeam (2020). *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA.

Sawusch, J. R. (1976). Selective adaptation effects on end-point stimuli in a speech series. *Perception & Psychophysics*, 20(1):61–65.

Sawusch, J. R. and Newman, R. S. (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Perception & Psychophysics*, 62(2):285–300.

Shinn, P. C., Blumstein, S. E., and Jongman, A. (1985). Limitations of context conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics*, 38(5):397–407.

Steffman, J. (2019). Intonational structure mediates speech rate normalization in the perception of segmental categories. *Journal of Phonetics*, 74:114–129.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5):1074–1095.

Toscano, J. C. and McMurray, B. (2010). Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics. *Cognitive Science*, 34(3):434–464.

Toscano, J. C. and McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, 74(6):1284–1301.

Wade, T. and Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics*, 67(6):939–950.

Welch, T. E., Sawusch, J. R., and Dent, M. L. (2009). Effects of syllable-final segment duration on the identification of synthetic speech continua by birds and humans. *The Journal of the Acoustical Society of America*, 126(5):2779–2787.

Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics*, 18:3–35.

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, New York.